Understanding how DNA methylation patterns at enhancers record cellular histories

By

Tim Scott

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December 16th, 2023

Nashville, Tennessee

Approved:

Emily Hodges, PhD, Advisor

Lea K Davis, PhD, Committee Chair

Eric Gamazon, PhD

Mary Philip, PhD

Melinda Aldrich, PhD

DEDICATION

To my incredible family.

To my mom, Anne, who has been the most supportive, encouraging, wonderful person I know.

To my dad, Terry, who has been both a supporting father and a tremendous role model,

To my wife, Brittany, who has been through most of this journey with me,

To my pups, Johnny and Winnie, friends and family forever.

ACKNOWLEDGEMENTS

I would not have been able to complete this work without the financial support of the Vanderbilt Human Genetics Program, Vanderbilt Genetics Institute, the NIH T32 training grant: Training Program on Genetic Variation and Human Phenotypes, and [Emily research grants (DoD, RO1, ACS)]. I would also like to thank my committee: Emily Hodges, Lea Davis, Eric Gamazon, Melinda Aldrich, and Mary Philip, who provided me with both scientific and professional guidance. Their insight helped to shape my dissertation into something I am proud of.

I would also like to especially recognize Emily Hodges, for taking me into her lab family, while encouraging me, supporting my goals, and facilitating my journey at each step. Without her patience and understanding, I wouldn't be here now. From relearning programming to discussing complex topics in half-sentences to Taco Bell runs, I'm glad IGP allowed me to find Emily as a mentor, and for Emily to be as dedicated to mentoring as she is.

I would also like to highlight Rosalind Johnson, a uniquely amazing, patient, understanding person who has been infinitely supportive, even if I don't see her every day. Her work and dedication to the program and its students is felt every day. She is another person, without whom, I believe my path would be very different.

Kathy Fries gave me the passion for genetics. I must also acknowledge my high school AP Biology teacher, Ms. Fries. Before her class, I disliked biology beginning in the 7th grade. I wanted to be a mechanical engineer or economist. However, in her notoriously difficult class, I was astounded at the amount of complexity, logic, interdisciplinary science, and interesting questions involved with biology. I would have never applied for undergraduate schools, focused on genetics programs, if it weren't for the inspiration from Ms. Fries.

Thank you Dr. Gerald Willing for taking in a 15-year-old high school student to do independent work on photovoltaic solar cells. There, I learned that science is not a linear journey and the fun that can be had in innovative techniques.

I must thank the lab of Dr. Connie Mulligan, especially Dr. Aida Miro, whose immense patience and passion for teaching introduced me to population genetics and bench work. There, I gained an immense appreciation for the amount of work, dedication, and ingenuity that occurs behind the scenes in a project. All of my undergraduate research successes are directly attributable to the support and knowledge I received from Aida.

The members of the Hodges lab are among the integral people in my journey to a PhD. Kelly Barnett, Johnathan Attalla, Tyler Hansen, Lindsey Guerin, Verda Miranda, Jessica Day, Kritika Singh, and Adam Miranda all contributed to the enjoyment of day-to-day lab work and to a positive lab environment. Between lots of jokes, random questions, troubleshooting, and stress-reducing outings, I'm very thankful for the faces who have been on this journey with me--if nothing else than to remind me that good science is a team effort, and I'm not alone.

I would also like to highlight my most diligent and long-term supporters: my parents. The unwavering support I have received provided a foundation from which I was able to explore the many opportunities I was afforded. Without them, I could not be here, and I would not be proud to be the student, friend, lab mate, and son, I am now. I will forever be thankful for the values, knowledge, opportunity, and support I have received from both my parents: Anne and Terry.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Definition and characterization of DNA methylation

DNA methylation (DNAme) is a chemical modification marked by the addition of a methyl group to the C-5 position of cytosine (5mC) within the CpG dinucleotide context. It is deposited by methylation writers, notably DNA methyltransferases (DNMTs). Methylation is stably maintained by DNMT1, a maintenance DNMT that preferentially methylates hemi-methylated DNA templates during DNA replication, while *de novo* methylation is handled by DNMT3a and DNMT3b. DNAme is also a heritable epigenetic mark, where methylomes can be faithfully reproduced through mitosis and cell expansion. Because DNAme changes are associated with changes in gene expression in the absence of alterations to the underlying DNA sequence, the modification is considered “epigenetic". Methylation has been studied in numerous contexts including X-chromosome inactivation (1), imprinting (2, 3), and transgene silencing. Here, I focus on the role of DNAme at gene enhancers in reinforcing tissue and cell identity.

Targeted mutagenesis of DNA methyltransferases—DNMT3a and DNMT3b—in mice results in embryonic lethality (4), demonstrating that methylation is essential for proper mammalian development. DNMT deficiency has also been shown to limit the ability of mouse embryonic stem cells to faithfully differentiate (5), indicating an important relationship between DNA methylation maintenance and cell specification. Additionally, in human embryonic stem cells, the deletion or inhibition of DNMT1 results in rapid cell death (6),

emphasizing the importance of DNA methylation maintenance in early cell development. CpGs may also be demethylated through a series of chemical transitions initialized by the ten eleven translocation (TET) proteins which function as 5-methylcytosine dioxygenases (7-9); this process involves a stepwise oxidation process that produces intermediates: 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. Like the methylation writers, the TET protein family also shows importance to developmental fidelity. Studies have shown that *TET2* gene deletions or loss-of-function mutations are common in numerous cancers, including myelodysplastic syndromes (~ 20%), acute myeloid leukemia (~ 12-25%), and chronic myelomonocytic leukemia (~ 20-40%) (10, 11). These studies emphasize the importance of DNA methylation mechanisms for driving proper cellular development.

While most of the twenty-eight million CpG sites genome-wide are stably methylated (~ 80-90% CpGs methylated) (12-14), scattered genomic regions of low methylation are enriched for transcriptional enhancers and promoters. CpG sites are generally depleted in mammalian genomes compared to other dinucleotides. This is due to the passive deamination of 5mC to thymine (15); however, areas of the genome that frequently lack methylation avoid this transition mutation and show lower rates of CpG depletion. Consequently, this leads to regions with higher frequencies of CpG dinucleotides than expected, called CpG islands, often coinciding with gene promoters and other regulatory elements (16, 17).

While CpG island definitions vary, they describe regions that feature high CpG density as a result of largely being hypomethylated (18, 19). CpG islands became a common methylation

annotation of study in early methylation work (12, 20-22), leading to increased knowledge of subsets of DNA regions of low methylation, or hypomethylation. CpG islands are associated with increased transcriptional permissiveness. They feature a lower concentration of nucleosomes associated with increased chromatin accessibility (23), yielding a transcriptionally permissive state that allows transcription factors, co-activators, and other proteins to interact with the underlying DNA sequence. Roughly 40% of transcription start sites coincide with a CpG island (24), while about 60% of CpG islands are found at promoters. Some transcription factor motifs are also enriched for CpG dinucleotides, implying methylation may play a role in some transcription factor binding (25), and thus, may impact gene transcriptional regulation at regulatory regions. Studies have shown some transcription factors to be methyl-sensitive (e.g. JUND, CREB1, NRF1 and CTCF) (26, 27), where methylation prohibits binding of the underlying sequence; other transcription factors have been revealed to have higher affinity for methylated CpGs (e.g. MBD1 and MBD2) (28, 29), and still, others are indifferent to methylation (e.g. YY1) (30). CpG islands are also associated with histone modifications indicative of transcriptionally permissive chromatin (23, 31). While CpG islands became a prominent annotation for HMRs, we now understand they represent only a subset of all HMRs across the genome (32, 33). A high density of CpG dinucleotides is integral to the CpG island definition. Dynamically methylated enhancers generally do not feature the CpG density to qualify for the heuristic criteria required for CpG island delineation; consequently, CpG islands enrich for promoters, and enhancer regions are largely ignored by this genomic annotation. While CpG islands persisted as a major annotation in early studies to understand DNA hypomethylation, in this

thesis, we analyze non-coding HMRs to better understand DNA hypomethylation at enhancers.

DNA methylomes between cell types are highly discriminatory of distinct cell states, marked by differences in cell-specific HMRs. While promoters are often stably hypomethylated, the most dynamically hypomethylated regions are distal from transcription start sites and enriched for active enhancers (14, 34-36). Differentially methylated regions are sufficient to distinguish cell types and recapitulate developmental relationships (13, 32, 33), indicating that methylation changes in the non-promoter context are associated with the gene regulatory events that reinforce and define cell identities. These observations also underscore the specificity of methylomes across cell types.

Historical perspective on DNA methylation and gene expression

DNAme has long been thought to modulate gene expression (37, 38), though the exact physical and temporal relationship between methylation and transcriptional permissiveness is poorly understood. While early studies prioritized the methylation of promoters as a mark of transcriptional silencing, results from various groups were incongruent regarding the strict idea of methylation as a silencing epigenetic mark. Early methylation studies using endogenous β-globin DNA suggested that DNAme was inhibitory to gene expression (39-41). Busslinger, et al. used an M13 (engineered to be hemimethylated) vector containing a β-globin gene transfected into mouse L-cells (a fibroblast cell line) to test the effects of methylation at different portions of a gene on transcription. Results showed that gene body methylation was permissive of transcription. However, methylation around the 5' region of

the gene led to a disruption of β-globin transcription, indicating the potential for promoter methylation to regulate gene expression. Another study using similar cloning techniques in mouse *Ltk*- cells found that unmethylated sequences introduced into a cell would integrate into the DNA in a DNase-I sensitive manner, indicative of open chromatin and a gene transcriptionally permissive state (42). In contrast, DNA sequences that were fully methylated lacked DNase-I sensitivity upon transfection, leading to the conjecture that methylated DNA induces inactive DNA structures.

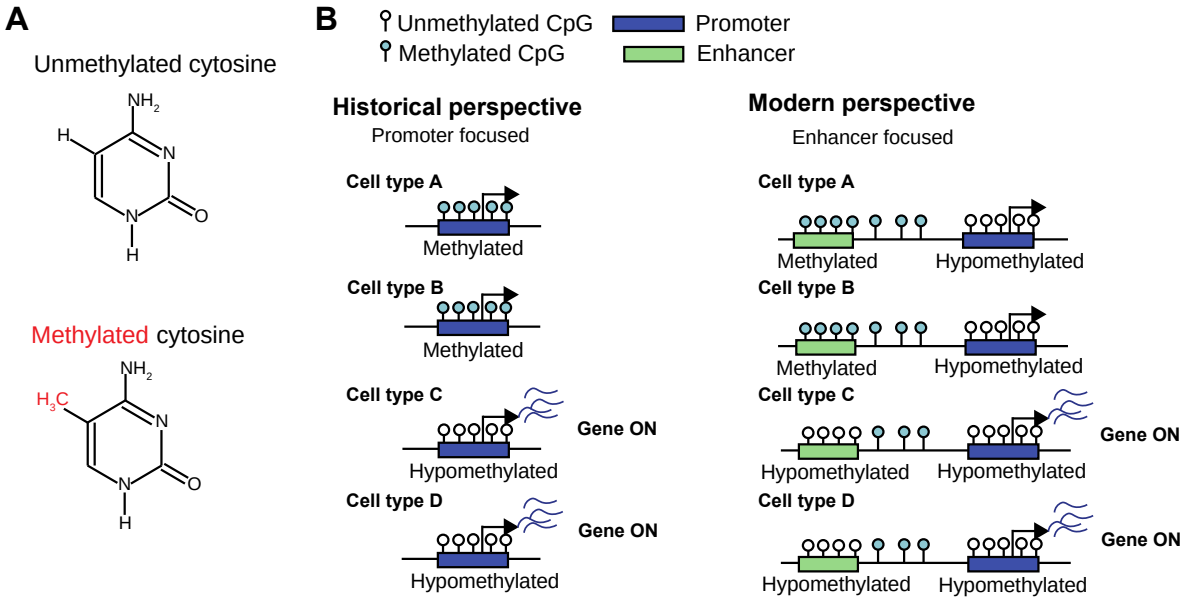
Around the same time, other researchers expanded beyond the mouse fibroblast L-cells to *Ltk*- cells (mononucleated myogenic cells that lack the leukocyte tyrosine kinase receptor) (43), which do not express endogenous α-actin. Researchers transfected methylated α-actin DNA into *Ltk*- cells as well as cells from the myogenic mouse *LB* line, which can be induced to express α-actin. In this comparison, methylation was sufficient to silence the expression of α-actin only in the fibroblasts. Interestingly, in the myogenic cells, methylation patterns of introduced genes included areas of demethylation that mimicked the pattern typically found in myoblasts *in vivo*. While the methylated promoter of α-actin prevented expression in fibroblasts, the methylated promoter did not affect transcription in myogenic cells; this suggests the methylation status is interpreted differently in distinct cellular contexts. These observations also suggested that methylation is not exclusively inhibitory (43).

A subsequent study investigated methylation of the IgG κ-gene in pre-B cells, finding similar conclusions. The IgG κ-gene can be transcriptionally induced with administration of lipopolysaccharide (LPS), where activation occurs although methylation changes in the pre-

B cells are absent over many generations (44). Thus, gene activation can be achieved independent of methylation changes. This suggested that methylation is not always, itself, inhibitory to expression and may otherwise function in a context-dependent manner with respect to transcriptional silencing.

Work measuring transcription and methylation of pluripotency genes in embryonic stem cell differentiation time courses suggested that methylation stably silences associated genes (45). More recent work has benefited from whole-genome methylation profiling in conjunction with RNA-seq data across multiple cell types, revealing that promoter methylation is largely invariable regardless of gene transcriptional status (46-48). Whole-genome bisulfite sequencing allowed for a deeper, genome-wide look at genome DNA methylation; results show that transcription is not strongly linked to promoter methylation as promoters generally feature low methylation across cell types regardless of transcriptional status (32, 34). It is important to note that subsets of promoters do undergo methylation associated with gene silencing during cell development. DNA methylation is associated with strong repression of parental alleles of imprinted genes (e.g. *H19* and *Igf2*) (49). Also, the promoters for Oct4 and Nanog, quintessential transcription factors for pluripotency, have been observed to become methylated during cell differentiation (13). While subsets of promoters undergo developmental methylation associated with gene silencing, promoter methylation is a poor predictor of transcriptional activity. Meanwhile, the non-coding genome (outside of promoters) shows dynamic methylation changes, suggesting methylation at enhancers and other non-coding regulatory regions is more correlated with transcriptional status (**Fig. 1**).

Figure 1. DNA methylation and transcriptional control paradigms.
(A) Diagram of the nucleic acid cytosine. The upper diagram shows a canonical cytosine. The lower graphic depicts a methylated cytosine; the additional methyl group is shown in red. (B) Diagram comparing the historical and modern perspective of the effect of enhancer/promoter methylation on transcriptional expression control. The historical perspective prioritized the methylation at promoters as a sign of transcriptional silencing. In comparison, the modern perspective prioritizes enhancer hypomethylation as an indicator of expression activity, as promoter hypomethylation is common regardless of transcriptional status.



Nonetheless, early studies in DNA methylation suggested a highly correlative relationship between promoter methylation and gene expression. Early studies involving the measurement of methylation and transcriptional status of genes revealed that transcriptional activation could precede the demethylation of regulatory elements (50, 51). While these studies suggest that DNA methylation is not always prohibitive of transcription, the notion remained that gene activation involved loss of DNA methylation at some point in developmental time. However, the timing and coordination of these two events remain poorly understood for decades. Thus, there remains an uncertain distinction as to whether DNA methylation is instructive or reflective of gene regulation.

Cross-tissue DNA methylation patterns

DNA methylation is highly conserved across species as well as tissues within a species (47), where methylation patterns are common among conserved sequences. Across the genome, methylomes are largely stable, as 85-90% of CpGs are constitutively methylated (13, 52, 53). The other variably methylated regions distinguish not only cell types but also lineages (14, 35), indicating that methylation patterns along differentiation pathways are stable and specific. A study by Eckhardt, et al. featured data from the Human Epigenome Project from diverse cell types including heart muscle, liver, skeletal muscle, sperm, fibroblasts, keratinocytes, melanocytes, and placenta (47). The sample collection also includes two very closely related hematopoietic blood cell types in CD4 lymphocytes and CD8 lymphocytes, which are both subtypes of T cells. The researchers assayed methylation at CpGs in chromosomes 6, 20, and 22 using targeted bisulfite sequencing in conjunction with ABI3730 capillary sequencing and found regions that were uniquely hypomethylated in subsets of cells. For example, the study highlights CpG islands that are specifically hypomethylated in only the lymphocytes but not other cell types. Results show additional subsets that uniquely define fibroblasts, keratinocytes, and melanocytes, underscoring the lineage and cell specificity of methylation patterns. This also suggests that methylation changes may be established at developmental stages that are then maintained as cells specify into various differentiated cell types, marking distinct lineages.

In another study comparing the methylomes of tissue- and cancer-specific CpG island shores, 2 kb regions that flank CpG islands and feature lower CpG-density than CpG islands (19, 54), this subset of hypomethylated regions were capable of distinguishing three cell types:

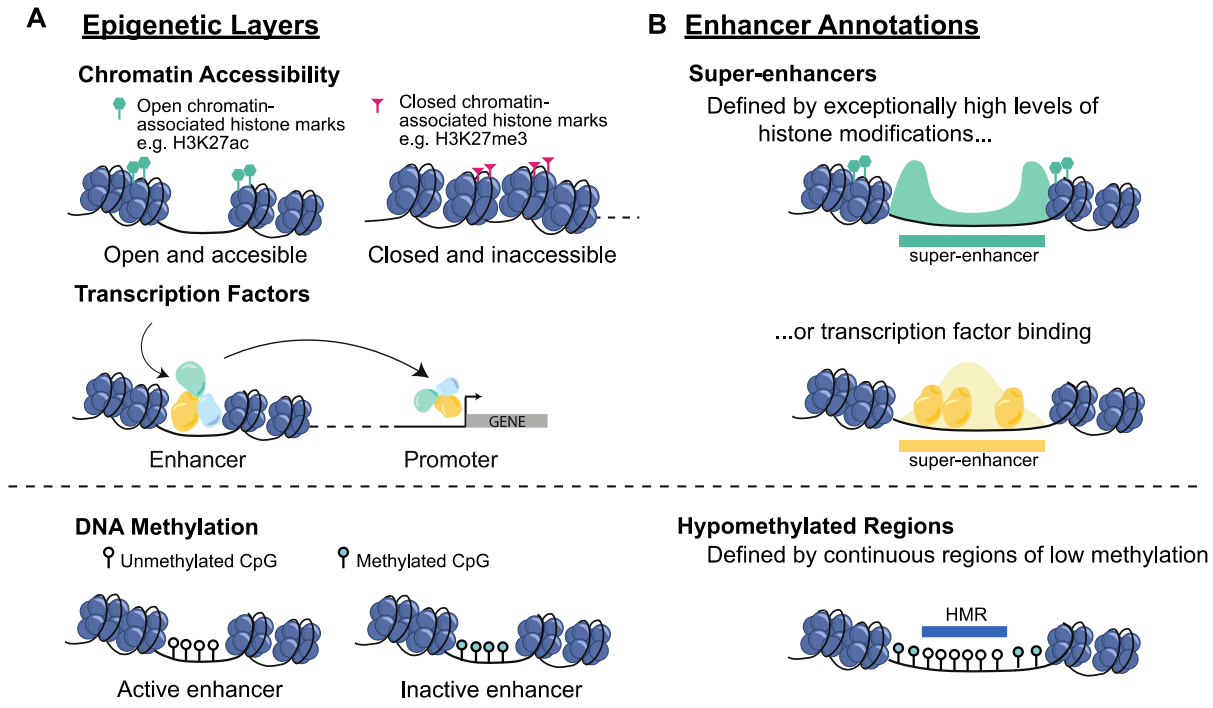
induced pluripotent stem cells (iPSC), embryonic stem cells, and fibroblasts, emphasizing the specificity of the methylome along developmental timepoints (55). The study further identified 4,401 differentially methylated regions between iPSCs and fibroblasts; using unsupervised clustering, the methylation values at these regions across samples were sufficient to fully distinguish normal brain, spleen, and liver samples. Using the same methodology, the methylation status at these 4,401 differentially methylated regions also largely delineated colorectal cancer from matched normal colonic mucosa, further identifying not only different cell types but also healthy cell states. While this collection of DMRs is somewhat small, the results emphasize the specificity of subsets of variably methylated regions. These limited subsets of HMRs (e.g. CpG islands and small DMR subsets) provided insight into the functional characterization of DNA methylation, but also presented limitations. Because earlier methylation studies relied upon locus-specific techniques (e.g. PCR) to study specific hypomethylated regions, genomic regions of interest were often limited to promoters, given the popularity of CpG islands. However, now whole-genome techniques have provided insight into non-coding HMRs as technology and analytic approaches have been refined.

As DNA methylation has been studied for decades, methodology for measuring CpG methylation has been diverse and improved over time, leading to advancements in both breadth and resolution. Early studies used a variety of methods to look at whole-genome levels of cytosine to methylcytosine, including a method involving hydrolyzation of the sample into nucleotides before employing chromatography to isolate fractions measured by relative UV absorbance (56, 57). Other assays relied on enzymatic digestion; briefly, by using

combinations of methyl-sensitive (i.e. HpaI can cut at unmethylated sites) and methyl-insensitive restriction enzymes (e.g. MspI) in conjunction with PCR, methylated and unmethylated fractions can be measured. The introduction of bisulfite conversion and sequencing allowed higher base pair resolution (58, 59). Bisulfite treatment of DNA results in the deamination of unmethylated cytosines to uracil, which is read as thymine after sequencing. Earlier methods utilizing Sanger sequencing or microarrays relied upon knowledge of a targeted region and allowed for locus-specific observations. The advent of next-generation sequencing provided an accessible method to assay the whole genome. This allows measurement of all CpG sites in a more unbiased manner. With the depth and resolution afforded by next-generation sequencing, we can better understand the genome-wide methylation trends.

Comparing whole-genome bisulfite sequencing data, the regions with the most dynamic methylation differences between cell types are found to be enriched in the non-coding genome (32). This is consistent with the idea that the dynamic epigenetic processes that control cell identity are controlled by the non-coding genome (60). Previous studies have highlighted the enrichment of active enhancers in non-coding HMRs (32, 34). Cell-specific HMRs show enrichment for enhancer-associated marks including histone modifications indicative of open chromatin (H3K27ac and H3K4me1), DNase I hypersensitivity, transcription factor ChIP-seq signal, and transcription factor motifs—all commonly utilized marks of active enhancers (**Fig. 2**). Additionally, regions that bind transcription factors show reduced levels of methylation (60-62), supporting the idea that dynamic non-coding HMRs are enriched in enhancers that impact cell identity.

Figure 2. Enhancer marks and annotations.
(A) Diagram of multiple epigenetic layers. Chromatin accessibility is largely influenced by and identifiable through numerous chemical histone modification. Transcription factors physically interact with the underlying DNA sequence and other co-factors to promote promoter-enhancer interactions. DNA methylation is the default status of the genome. Continuous regions of low methylation are indicative of promoters or non-coding regulatory elements. (B) Chromatin immunoprecipitation sequencing (ChIP-seq) can be used to identify histone modifications or transcription factor binding. Exceptional levels of ChIP-seq values for histone modifications (often H3K27ac) or specific transcription factors can be used to define super-enhancers (top). The bimodal nature of DNA methylation reveals continuous regions of low methylated CpG sites that define hypomethylated regions with well-defined boundaries.



Multi-unit enhancer annotations

Early locus-specific studies identified regulatory segments of the genome that could control transcription of a target gene. These regions, labeled locus control regions (LCRs), were first described in the β-globin locus. While a 5 kb segment of the β-globin gene was transcribed in erythroleukemia cell lines, it was not observed to promote strong or detectable expression

in transgenic mice (63, 64). It was observed that the deletion of a region upstream of the β-globin gene was common in β-thalassemia cases (65-67), where the absence of this regions resulted in a lack of chromatin accessibility at the locus. The addition of the LCR, located between 6 and 22 kb upstream of the β-globin gene, yielded expression of the β-globin gene segment at near-endogenous levels, suggesting that the LCR, a genomic region distal from the actual gene, was necessary for proper gene expression (68). The β-globin LCR consists of five distinct DNase-I hypersensitivity sites (69). The first four are present in only erythroid cells, while the fifth is formed in multiple other lineages. Targeted analysis using fragments of the LCR identified context-dependent enhancer activity from DNase hypersensitivity sites (DHSs) 2, 3, and 4. DHSs represent regions of open chromatin, which are measured by levels of DNase digestion, indicative of chromatin accessibility. The activity of the LCR is erythroid cell type-specific, highlighting the ability for individual units of enhancer clusters to behave in an independent and context-specific manner.

Similarly, investigation of the *cis*-regulatory landscape surrounding pancreatic islet genes highlighted the role of clustered enhancers in tissue-specific gene regulation (70). One thousand pancreatic islet-specific genes and ubiquitously transcribed genes were compared with transcription factor ChIP-seq and chromatin accessibility data. Most genes with islet-specific gene expression showed a high surrounding density of chromatin accessibility sites, with an average of three clustered enhancers compared to a single enhancer around ubiquitously expressed genes. These spatially related enhancers were also enriched for Type 2 Diabetes risk-associated variants, highlighting the functional roles of the clustered

enhancer regions in tissue-specific biology. This context-specific example highlights the relationship between tissue-specific gene regulation and clustered regulatory elements.

Super-enhancers are defined as regions of the genome that show exceptionally higher levels of ChIP-seq signal, generally measuring either transcription factor binding (e.g. Med1, Klf4, or Esrrb) or H3K27ac, a mark of accessible chromatin (71). More specifically, they are identified by ranking all ChIP-seq regions by total background-subtracted ChIP-seq signal on the *x*-axis and plotting the total background-subtracted ChIP-seq signal (reads per million per bp) on the *y*-axis; the method then identifies the point on the *x*-axis where a tangent line to the resultant curve has a slope of 1, where any region to the right of that point (higher ChIP-seq signal) is defined as a super-enhancer. The process of defining super-enhancers also includes a stitching step before ranking that involves combining regions within 12.5 kb end-to-end into a continuous region. While not all super-enhancers consist of multiple individual units, they are commonly referred to as collections of clustered enhancers (56% consist of multiple units) (72). They are enriched for transcription factor motifs and are physically near genes that regulate developmental specification and reinforce cell identity. Super-enhancers are also enriched for disease-associated variants and eQTLs, suggesting that clustered enhancers are significant for linking genome to phenome. We find that clustered HMRs are especially enriched for variants linked to cell-specific phenotypes.

Stretch enhancers, another annotation that includes clustered enhancer regions, are genomic regions defined by ChromHMM enhancers states (regions of the genome that are labeled as likely enhancers by a Hidden Markov model trained to recognize genome states based on ChIP-seq data targeting histone modifications [e.g. H3K27ac and H3K4me1] and

CTCF) are continuous and have a length that exceeds 3,000 base pairs. Like super-enhancers, they are also enriched for specific transcription factors motifs and GWAS SNPs (73, 74).

While numerous chromatin-based methodologies have revealed clustered enhancer loci, their identification varies depending on the ChIP-seq target used to find them. In recent work in our lab, we have observed that HMRs also cluster together more often than expected by random chance. HMRs can be linked together end-to-end (maximum inter-HMR distance of 6 kb) between individual HMRs to compose clusters of HMRs. We find HMR clusters to be a larger genomic annotation with more regions that partially encompasses other clustered enhancer annotations. We compared HMRs with super-enhancers, finding a small congruent subset; however, the majority of clustered HMRs do not overlap with SEs—currently a prominent annotation used to describe clustered enhancers. Additionally, both our data and that of others reveals that HMR clusters exist outside of open chromatin; this suggests that clustered HMRs are not only more permissive than clustered enhancers defined by chromatin accessibility, but may record genome-phenome information not retained by chromatin accessibility marks (e.g. DNase-seq or ChIP-seq data for histone modifications).

Locus-specific studies have revealed more complex dynamics within clusters of enhancers. A study of the STAT5-driven *WAP* enhancer in mammary tissue showed an interplay between three individual enhancers within a super-enhancer (75). Targeted mutations in transcription factor binding sites within each individual element showed that distinct enhancers had differential effects with regards to transcriptional deficiencies. The inhibition of individual or combinations of enhancers revealed differential impact on transcriptional

regulation, indicating that not all enhancers in a cluster are equal in a cell-specific context. Furthermore, the same paper showed that while all super-enhancers in the study were defined by similar STAT5 binding, only half were associated with highly expressed genes associated with the induction of STAT5 during pregnancy. In fact, the authors report an associated transcriptional range that spans over four magnitudes, suggesting expression at associated genes is widely variable and can include low expression. This suggests that all super-enhancers are not characterized by the classical assumption of strong transcriptional effects, and that clustering enhancers may have other characteristics that extend beyond the super-enhancer annotation. Super-enhancers have a specific definition reliant upon "exceptional" levels of ChIP-seq signal, where not all super-enhancers comprise individual component elements. Rather, we find that clustering is more prevalent in methylome data than in super-enhancer defined regulatory elements (76).

Another locus-specific study in mouse used a combination of p300 ChIP-seq binding data and a reporter assay to identify a grouping of three enhancers associated with *Sox2* (77), which is a key transcription factor attributed with maintaining pluripotency. Using heterozygous deletion of individual enhancers, the study showed differential implications for transcription of *Sox2*. Deletion of some enhancers reportedly did not affect expression levels, while deletion of others significantly reduced mRNA and protein levels of Sox2. This reinforces the idea that the individual elements within a cluster do not share the same enhancer functions in a given cellular context. The deletion of these elements resulted in aberrations to cell colony morphology, gene expression, as well as the ability to differentiate into embryoid bodies, highlighting the potential significance of enhancer clusters to faithful

differentiation and proper cell identity. Super-enhancers are defined by ChIP-seq signal of specific enhancer-associated marks, including histone modifications and transcription factor binding. However, DNA methylation was largely ignored as an enhancer-associated epigenetic mark for the study of clustering enhancers. Here, we have revealed that HMR clusters exist outside of chromatin accessible regions and describe a previously underappreciated clustered enhancer mark.

Using electronic health records to study HMR function

Electronic health records (EHRs) have been adopted as a strategy to obtain large numbers of cases and controls for diseases and phenotypes of research interest. EHRs contain medical data on patients in a hospital system over time and may be paired with genotyping data providing a rich resource for clinical epidemiological and genetic studies. Records may include patient billing codes, procedural codes, medication history, clinical notes, and other demographics. The availability of large quantities of disease cases accompanied by genotypic data allows for the testing of statistical associations between genetic variants and phenotypes. BioVU, the biobank effort at Vanderbilt University, provides the ability to test for genetic associations with both diseases in a phenome-wide association study (PheWAS) as well as clinical lab values in a lab-wide association study (LabWAS) through logistic and linear regression (78-80).

Biobanks have started to proliferate across the world, capturing patient populations with diverse multi-ancestral backgrounds and environmental exposures. The sample sizes provided by modern, continuously expanding biobanks provides the opportunity to look for

statistical trends between genetics and phenotypes. Researchers can statistically associate the presence of an alternative allele at a SNP with clinical lab values or disease traits. Major biobank efforts include both academic and private institutions, including FinnGen, Biobank Japan, UCLA Precision Health Biobank, Michigan Genomics Initiative, and BioVU (81). Sample sizes at individual sites (e.g. ~120k in BioVU and ~500k in the UK Biobank) have provided the ability to statistically test for variant-trait associations in well-powered phenotypes. Differences in ascertainment strategies (i.e. some of these efforts involve participant acquisition through population-based health programs while others are ascertained from a health center context), sequencing methods, and phenotypes represented in the EHRs make meta analyses and comparisons across biobanks more difficult. Nonetheless, with efforts such as the Global Biobank meta-analysis Initiative to standardize methods across global sites with the goal of combining sample sizes across broad backgrounds, biobanks offer the promise of revealing novel genetic and trait associations. Overall, biobanks have contributed to research for genetic associations with disease phenotypes, providing potentially actionable targets for genetic treatments.

Genome-wide association studies (GWAS), which utilize large amounts of individual genotypic data with EHRs to look for SNPs of interest in association with a clinical trait, have revealed that most disease-associated genetic variation is in the non-coding genome (82-84). GWAS variants are also enriched for expression quantitative loci (eQTLs), which are sequence variants that are statistically associated with altered expression of a gene (85, 86). This suggests that the most common disease-associated sequence variant does not affect the coding region, and putative protein function, but rather affects the expression of genes

through regulatory elements. As HMRs are enriched for non-coding enhancers (32), these observations indicate that HMRs feature trait-associated genetic variation (60). The use of GWAS SNPs may provide biological context for interpreting the function of non-coding HMRs. We find that HMRs established in developmentally distinct contexts tag enhancers relevant to the gene transcription regulatory needs at distinct developmental stages. This provides a framework for better understanding the role of DNA methylation as well as providing context for genetic variants within HMRs. However, it remains difficult to assign the exact role and potential gene targets of individual HMRs. Strategies to determine enhancer-gene relationships commonly rely on nearest neighbor approaches, where about half of gene assignments are incorrect (87); to accurately ascertain the proper gene target(s) would require time-consuming mechanistic studies with very low throughput. By aligning genetic variant-trait data with methylation data, we can link an HMR more directly to a biological function by utilizing directly overlapping sequence variants.

Currently, we can infer the function of non-coding putative enhancers by associating the regions to nearby genes, thought to be potential regulatory targets. We can then use knowledge of the genes to infer the function and spatiotemporal specificity of their regulatory elements with gene ontology; this strategy identifies enrichments of ontological terms that have been assigned to groups of genes by similar biological function. However, studies have shown that utilization of nearest neighbor methodologies for gene association are at best nearly 50% incorrect. Nonetheless, without the use of high-resolution chromatin capture techniques to measure the 3-dimensional relationship of the genome, nearest neighbor approaches tend to be the most reliable methods (87). This highlights the difficulty

in using traditional gene assignment methods. By comparison, we can identify genetic sequence variants directly overlapping HMRs that were directly associated with a disease state, thus linking the function more directly to the HMR sequence.

Stratified LD score regression (S-LDSC) is another method to associate sequence variants with clinical diseases and lab values (88, 89). The methodology attempts to quantify the enrichment of SNP-based additive genetic heritability—or the heritability captured by underlying genetic variants as opposed to environmental influences. The "partitioned" heritability method aims to quantify the amount of heritability captured by a subset of the genome (e.g. super-enhancers, open chromatin regions, promoters, or HMRs), normalized by the amount of SNPs included within that annotation ([% h^2]/[% SNPs]) (89). S-LDSC estimates heritability based on the concept that SNPs with high linkage disequilibrium (LD) to other SNPs are more likely to capture a causal genetic variant. For one SNP, the statistic from GWAS summary statistics should capture the total effect of SNPs in LD with that particular SNP. Thus, LD score (a measure of total LD for a SNP) is proportional to the X^2 statistic. If an annotation is enriched for heritability, then that category should contribute more to the X^2 statistic than another category with lower enrichment for heritability. The method attempts to find annotations where SNPs with high LD to that category also have higher X^2 statistics than SNPs with lower LD to that category. The model estimates effect sizes of SNPs linearly on the input annotation categories—in other words, the strategy asks if any category contributes more to the genetic heritability than other annotations.

S-LDSC is an attractive approach as it uses GWAS summary statistics to estimate LD scores within samples; this is much more feasible and computationally tractable than computing heritability estimates from individual patient-level data and genotypes. As summary statistics are widely available across various diseases and lab values, we are enabled to examine partitioned heritability estimates from diverse phenotypes of importance to the cell types we include in our analyses. In our own work, we have employed S-LDSC to assess partitioned heritability among our HMR groups defined by developmental specificity (e.g. within the hematopoietic lineage) or clustering.

Scope of Thesis

In this dissertation, I present my primary project to investigate the functional role of HMR patterns in defining cell histories. In Chapter II, I present the findings of my research focused on the patterns that arise from comparing whole-genome methylation profiling across diverse and highly related cell lineages. While previous studies have focused on limited pairwise differential methylation comparisons and locus-specific changes, we utilize whole-genome bisulfite sequencing to investigate enhancer HMR patterns both within and between cell types. By analyzing the correspondence of non-coding HMRs across diverse human cell types and tissues, we identify a hierarchical conservation of HMRs, enriched for stage-relevant enhancers, along cell differentiation trajectories. HMRs established at distinct developmental contexts capture scaling genetic heritability of cell-relevant complex traits, underling the power of HMR patterns to inform the function of the underlying DNA sequence.

We expand on these observations to show that HMRs accumulated through cell development are established near existing HMRs; this leads to the formation of HMR clusters. We show that HMR clusters are established near active genes important for cell identity, enrich for regulatory elements, and capture a disproportionate amount of partitioned genetic heritability relative to their unclustered counterparts. By comparing against super-enhancers defined by ChIP-seq signal for histone modifications or transcription factor binding, we find HMR clusters to be more pervasive, indicating that clustered enhancers may be underappreciated and suggesting a unique epigenetic role for DNA methylation. Collectively, these data reveal how DNA hypomethylation reflects previous and current genome function, providing genetically distinct epigenetic records of cell developmental states.

CHAPTER II

CROSS-TISSUE PATTERNS OF DNA HYPOMETHYLATION REVEAL GENETICALLY DISTINCT HISTORIES OF CELL DEVELOPMENT

BACKGROUND

Among the twenty-eight million CpG dinucleotides in the human genome, the majority (80-85%) of cytosines undergo constant DNA methylation (DNAme) in most cellular contexts (52, 53, 90-92). However, a subset of sites forms discrete regions containing stretches of CpGs that are not covalently modified by methylation and are thus considered "hypomethylated". The majority of these hypomethylated regions (HMRs) are non-coding and coincide with putative gene regulatory elements including promoters and enhancers (14, 34-36, 53).

DNAme has long been tied to transcriptional control; however, apart from a very small subset of developmentally regulated genes, promoters are stably hypomethylated and largely invariant across cell types, regardless of gene transcriptional status (47, 93-95). Thus, promoter HMRs poorly predict transcriptional programs that ultimately determine cellular phenotypes. By contrast, enhancer HMRs vary considerably between cell types, which results from their context-dependent demethylation (32, 33, 57, 96-101). While enhancer HMRs are more predictive of nearby gene activity than promoter HMRs (34), how these HMRs are established or maintained and their relationship to cell identity is not well understood.

We previously showed that non-coding HMRs represent an exclusive subset of chromatin accessible sites (34). More recently, we showed that, while HMRs correlate with chromatin accessibility and other indicators of permissive chromatin, the temporal dynamics of HMR formation is distinct from chromatin remodeling changes (99, 102). Importantly, HMRs can persist long after chromatin remodeling changes during cell fate transitions in terminally differentiating hematopoietic cells (99, 102). Similarly, in the mammary gland, gene regulatory changes during the first pregnancy result in

demethylation of pregnancy-responsive gene enhancers. The maintenance of these enhancer HMRs is long-lasting, even after pregnancy signals dissipate (103). These studies indicate that HMRs capture both active and previously active gene regulatory elements in a manner not reflected by other common enhancer-associated chromatin states.

Despite these observations, very few studies have considered the combinatorial and temporal significance of HMR patterns in a genome-wide manner across developmentally diverse datasets. For example, "super-enhancers", a class of enhancers that are defined by high levels of histone H3 lysine 27 acetylation (H3K27ac) and Mediator binding, are often comprised of multiple enhancers units (71, 104). Both selective and persistent hypomethylation of individual enhancer units within super-enhancers have been observed in mouse embryonic stem cells (ESCs) during exit from naïve pluripotency (62, 105). These combinations of HMR patterns suggest that coordinated hypomethylation of enhancers through cell fate transitions serves to uphold specific cellular states. Altogether, this argues that HMRs are established and maintained as a memory of gene regulatory activity; thus, consideration of how HMRs are shared within and between cell types may inform critical epigenetic patterns that secure cellular phenotypes. However, this hypothesis and its link to phenotypic outcomes remains to be tested across diverse tissues and developmental timepoints in a genome-wide manner.

Here, we performed a comparative analysis of whole-genome methylation data from diverse tissues representing distinct organ systems and developmental timepoints. Unlike previous studies that emphasize pairwise differential methylation or locus-specific changes during limited differentiation time courses, we comprehensively characterize HMR relationships both within and between cell types to understand the functional significance of combinatorial HMR patterns. By analyzing methylomes across diverse cell types, both distant and related, we show that hierarchical conservation of HMRs across tissues can identify enhancer HMRs established in developmentally

distinct contexts. We further demonstrate that HMRs established at distinct timepoints partition the genome in a way that is highly predictive of complex trait heritability, which highlights the significance of these HMR patterns to the underlying genome sequence. Ultimately, these data provide novel insights into how DNA hypo-methylation informs genome function by providing a map that traces the developmental histories underlying cellular states.

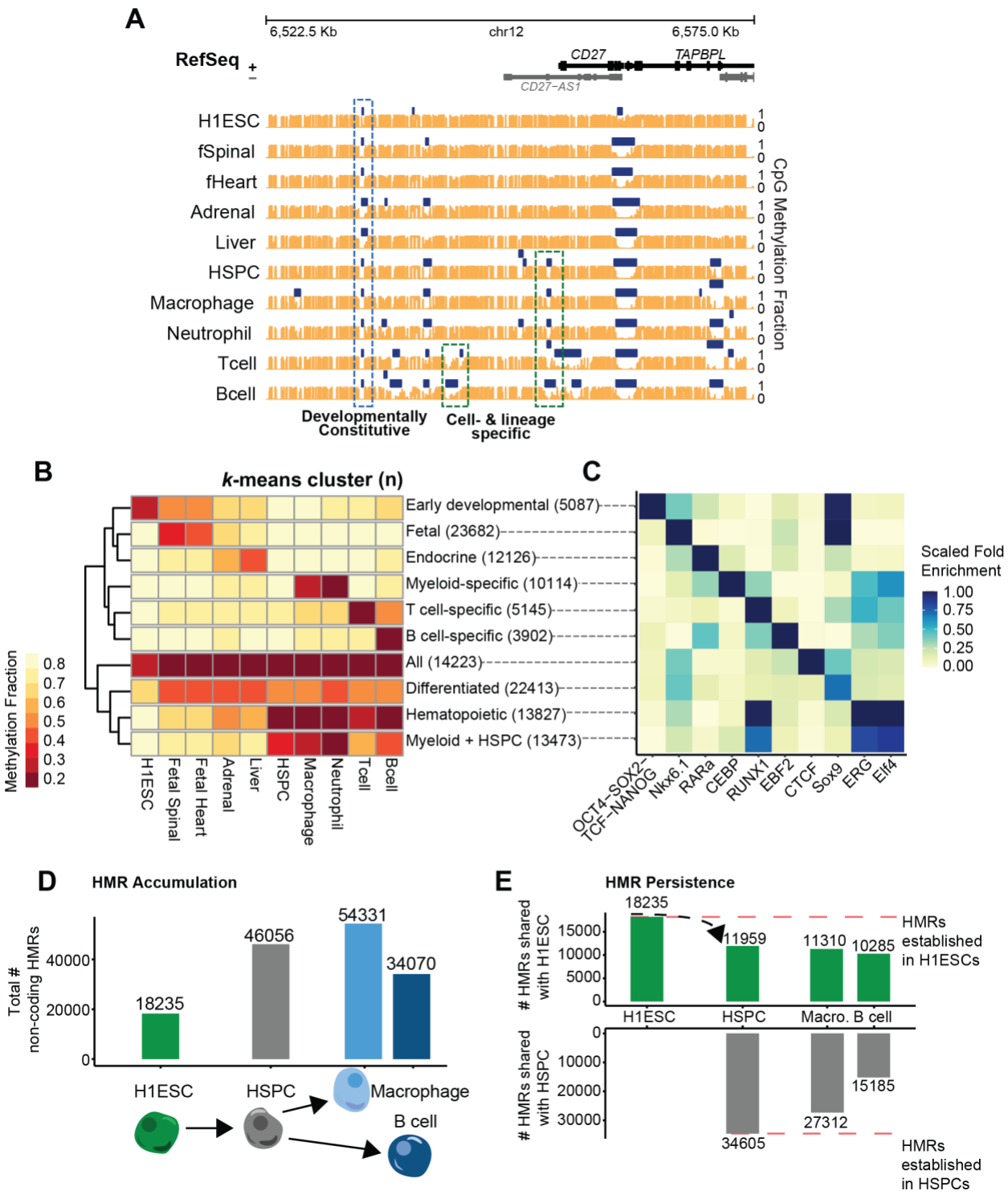
RESULTS

Shared HMR patterns among diverse cell types reveal common functional and developmental histories.

Studies aiming to understand the relationship between DNA methylation patterns and phenotypic outcomes have focused largely on individual differentially methylated regions without consideration of combinatorial changes that drive phenotypes. To understand the functional significance of complex HMR patterns, we determined the correspondence of HMRs across diverse human cell types and developmental timepoints. We hypothesized that shared HMR patterns among diverse cell types could reveal common functional and developmental histories. To illustrate this idea, genome browser tracks of methylation data are displayed for datasets representing diverse lineages and developmental timepoints at a B cell enhancer cluster upstream of the *CD27* gene (**Fig 3A**). This locus contains a group of HMRs with varying levels of HMR specificity that are surrounded by genes involved in lymphoid development and signaling including *CD27*, *LTBR*, and *TAPBPL*. A comparison of HMRs reveals different levels of both cell-type and lineage specificity, including HMRs conserved in all samples (developmentally constitutive); HMRs shared exclusively among lineage-related samples (e.g., hematopoietic cells); and HMRs present only in B cells. The lymphocyte-specific expression of *CD27* highlights a potentially important role for the combination of shared and cell specific HMRs observed at this locus.

Figure 3. Levels of HMR specificity recapitulate developmental relationships through accumulation and maintenance.

(A) Multiple alignment of WGBS methylation and HMR tracks across 10 cell types: H1 ESC, fetal spinal cord, fetal heart, adrenal gland, liver, hematopoietic stem and progenitor cells, neutrophil, macrophage, B cell, and T cell. Methylation tracks are represented by orange vertical bars showing methylation value per CpG site. Methylation fraction is calculated as the fraction of reads containing a methyl-C over the total number reads covering a CpG site. HMR calls are shown by dark blue horizontal bars. Developmentally constitutive, lineage-specific, and cell specific HMRs are highlighted by blue and green dotted bars, respectively. The *plotgardener* R package was used to generate the genome browser snapshot (106). (B) Heatmap of average methylation per HMR across cell types. Non-coding HMRs were *k*-means clustered based on their average CpG methylation values across 10 cell types represented in (A). A *k*-means of 10, assessed by the elbow method, was used to cluster HMRs into groups that are consistent with the biological relationships of their cell types. Groups are manually labeled to reflect their biological relationships. (C) The transcription factor (TF) motif enrichment of each *k*-means group reflects biological relationships captured in (B). Representative TFs were selected from the top significant hits ranked by natural log adjusted *p*-value for each *k*-means group. The top ranked TFs are shown unless the top TF(s) for that group were redundant; the second top ranked TF is shown for the group, "Myeloid + HSPC," and the third ranked TF is shown for the group, "Differentiated." Fold enrichment values are normalized from 0 to 1 across TFs. The background comparison file comprises HMRs across all represented cell types. **(**D) Bar graph of the total number HMRs for each cell type, arranged by developmental progression. (E) Bar graph measuring the presence of HMRs established in either H1 ESCs (*top; green)* or HSPCs (*bottom; grey*) in developmentally progressive cell types. The software *Bedtools intersect* was used to determine overlap between cell type HMR datasets using default settings (107). Overlap was defined as a 1bp minimum.


A
6,522.5 Kb
chr12
6,575.0 Kb
RefSeq
CD27
TAPBPL
CD27-AS1
H1ESC
fSpinal
fHeart
Adrenal
Liver
HSPC
Macrophage
Neutrophil
Tcell
Bcell
CpG Methylation Fraction
Developmentally Constitutive
Cell- & lineage specific
B
k-means cluster (n)
Early developmental (5087)
Fetal (23682)
Endocrine (12126)
Myeloid-specific (10114)
T cell-specific (5145)
B cell-specific (3902)
All (14223)
Differentiated (22413)
Hematopoietic (13827)
Myeloid + HSPC (13473)
Methylation Fraction
H1ESC
Fetal Spinal
Fetal Heart
Adrenal
Liver
HSPC
Macrophage
Neutrophil
Tcell
Bcell
C
Scaled Fold Enrichment
OCT4-SOX2-TCF-NANOG
Nkx6.1
RARa
CEBP
RUNX1
EBF2
CTCF
Sox9
ERG
Elf4
D
HMR Accumulation
Total # non-coding HMRs
18235
46056
54331
34070
H1ESC
HSPC
Macrophage
B cell
E
HMR Persistence
HMRs shared with H1ESC
18235
11959
11310
10285
HMRs established in H1ESCs
H1ESC
HSPC
Macro. B cell
HMRs shared with HSPC
34605
27312
15185
HMRs established in HSPCs

To investigate the extent to which these HMR patterns can be observed globally, we determined a set of high-confidence HMRs using publicly available whole genome bisulfite sequencing (WGBS) data from ten different cell types and tissues, including embryonic stem cells (H1 ESCs), hematopoietic stem & progenitor cells (HSPCs), fetal heart, fetal spinal cord, liver, adrenal gland, macrophages, neutrophils, T cells, and B cells (see Methods). As the resolution of HMR specificity is contingent on the quantity and interrelatedness of cell types included in the analysis, we maximized comparative potential by including datasets representing a diversity of organ systems and developmental stages.

HMRs were determined for each dataset using MethPipe (108, 109), which employs a computational model originally described in *Molaro et al. 2011* (33) to detect adjacent clustering of unmethylated CpG sites in the genome. Specifically, a 2-state hidden Markov model (HMM) with Beta-Binomial emission distributions allowed high and low methylation states to be trained separately on each individual WGBS dataset (108, 109). This modeling approach is robust to sequence coverage differences both within and between WGBS datasets. This is important given that sequence coverage is not uniformly distributed across the genome. Therefore, we required a minimum mean sequence read coverage of 10x at symmetric CpG sites for any HMR dataset to be included in our analysis. Of the ten datasets included, eight achieve CpG read coverage >25x, while the B and neutrophil cell datasets reach nearly 12x (**Table 1**) (108). This resulted in a total set of 126,104 unique non-coding HMRs with an average length of ~866 bp (**Fig 4**). Those HMRs spanning transcriptional start sites (TSS; -2000/+1000 bp) and exons were excluded from the analysis in order to focus on non-coding HMRs harboring putative enhancers (**Table 2**). By excluding the substantial number of constitutive HMRs overlapping gene promoters, we achieve better resolution to detect non-promoter HMR patterns that contribute to cellular states.

Figure 4. HMR lengths by cell type.
Density plot of HMR lengths (in bp) by cell type. The x-axis of the plot is visually limited to the range of 0 to 5000 bp for visibility.

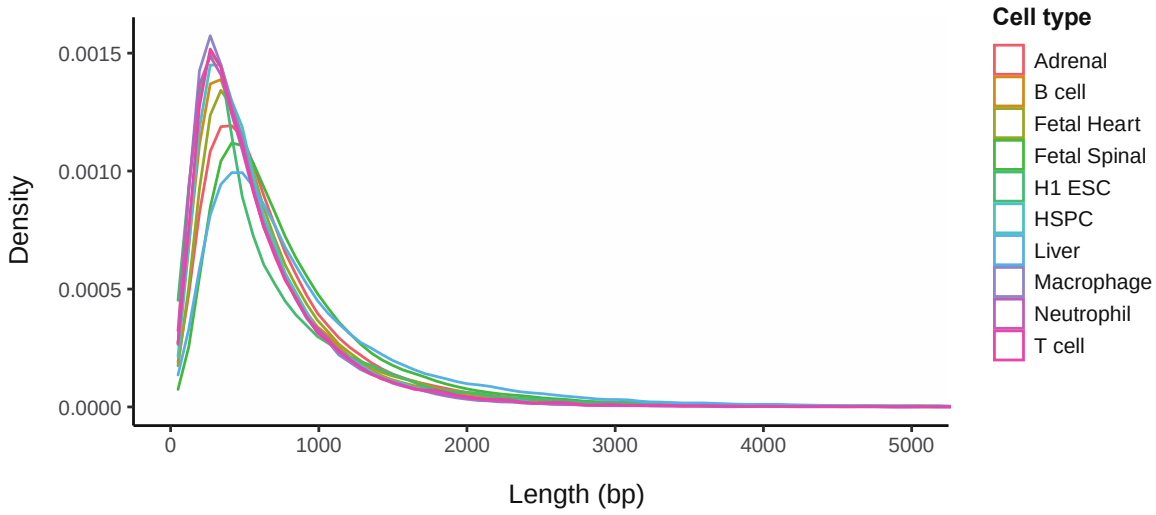


Table 1. Table of coverage values for WGBS datasets per cell type.

Cell type	Coverage	Download link
H1 ESC	25.933	http://smithdata.usc.edu/methbase/data/Lister-ESC-2009/Human_H1ESC/tracks_hg19/Human_H1ESC.hmr.bb
Fetal heart	37.134	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Fetal-Heart/tracks_hg19/Human_Fetal-Heart.hmr.bb
Fetal spinal cord	33.623	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Fetal-Spinal-Cord/tracks_hg19/Human_Fetal-Spinal-Cord.hmr.bb
Adrenal	71.558	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Adrenal-gland/tracks_hg19/Human_Adrenal-gland.hmr.bb
Liver	49.478	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Liver/tracks_hg19/Human_Liver.hmr.bb
HSPC	37.562	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_HSC/tracks_hg19/Human_HSC.hmr.bb
Macrophage	36.130	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Macrophage/tracks_hg19/Human_Macrophage.hmr.bb

Neutrophil	11.602	http://smithdata.usc.edu/methbase/data/Hodges-Human-2011/Human_BCell/tracks_hg19/Human_Neut.hmr.bb
B cell	11.855	http://smithdata.usc.edu/methbase/data/Hodges-Human-2011/Human_BCell/tracks_hg19/Human_BCell.hmr.bb
T cell	34.106	http://smithdata.usc.edu/methbase/data/Roadmap-Human-2015/Human_Tcell/tracks_hg19/Human_Tcell.hmr.bb

Table 2. Number of HMRs after preliminary filters.

	# HMRs Total Raw	# HMRs Post-50 bp filter	# HMRs Post-50 bp filter Post RefSeq TSS/Exon filter
H1 ESC	36359	35965	18235
Fetal Spinal	65130	65105	44390
Fetal Heart	64186	64122	43473
Adrenal gland	56655	56549	36610
Liver	58652	58559	38132
HSPC	67223	67069	46056
Macrophage	77058	76898	54331
Neutrophil	72120	71731	49103
T cell	51640	51539	32366
B cell	54998	54792	34070

Hierarchical clustering applied to these datasets was sufficient to recapitulate both related and distant cell type relationships, demonstrating the quality and specificity of HMR calls (**Fig 5**). Next, we utilized *k*-means clustering to group HMR methylation levels across the 10 different cell types and tissues in an unsupervised manner. We used the elbow method to determine an optimal number of *k*-means clusters (n=10**, Fig 6**).

Figure 5. Hierarchical clustering of HMRs by average methylation per cell type.
Dendrogram of average CpG methylation across HMRs per cell type. The input matrix used for the *k*-means clustering heatmap in Fig 3 was used for input to the R program, *ggdendro*. Distance was measured with the "euclidean" option, and hierarchical clustering was performed with the *ward.D2* method.

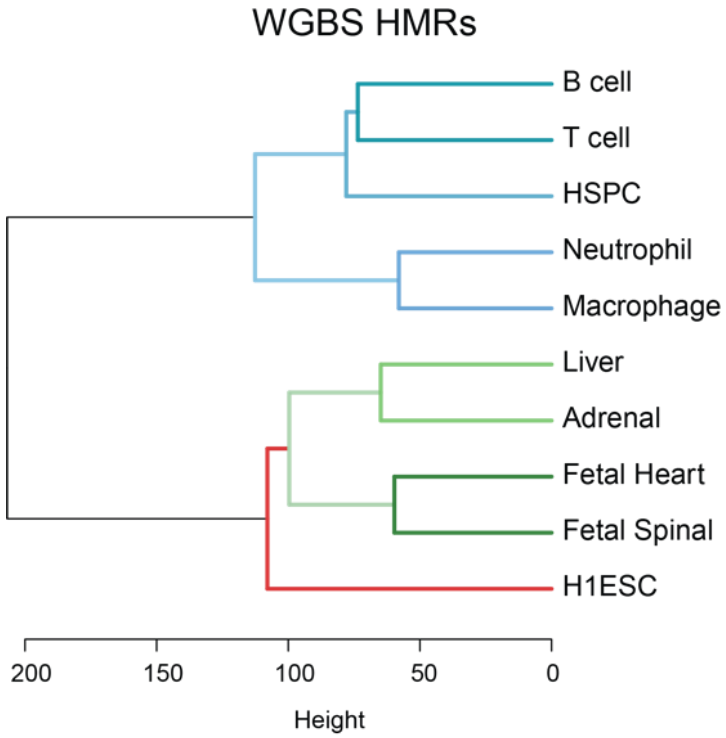
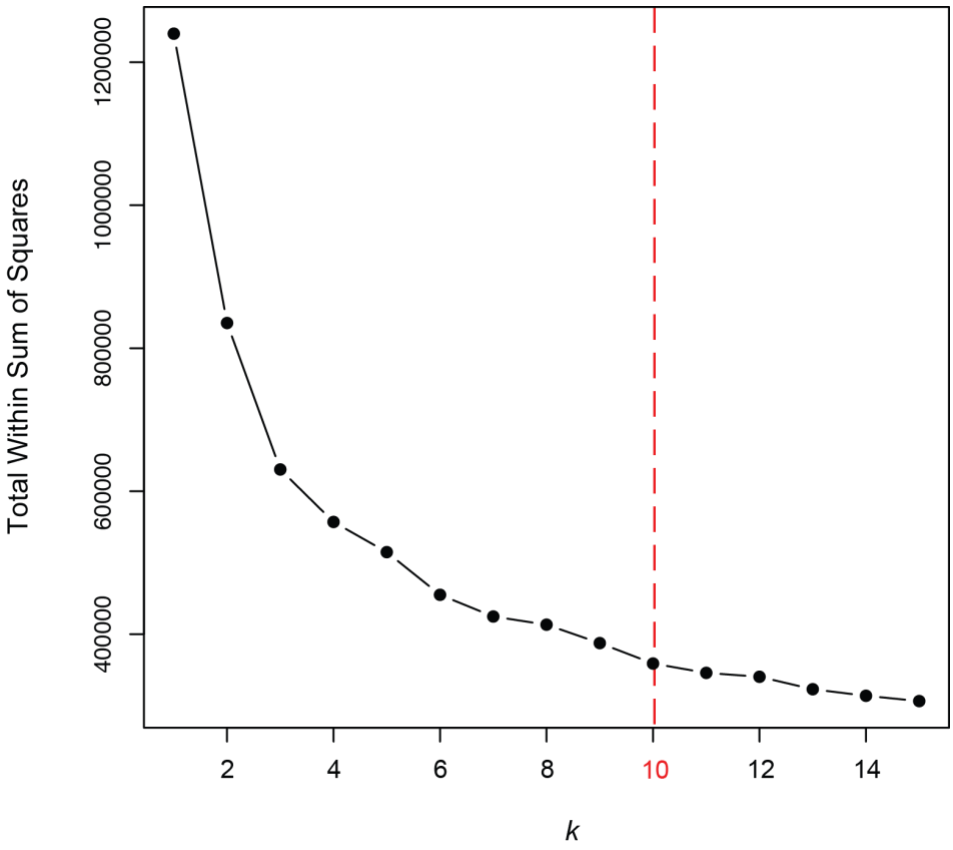


Figure 6. Dotplot of elbow method to determine appropriate number of *k*-means for methylation heatmap.
Figure displays within sum of squares estimates for clusters at each value of *k*-means group amount from 1 to 2. Estimates are derived from the *kmeans()* function in R.



The resultant heatmap revealed groups of HMRs highly stratified by both group function and developmental stage (**Fig 3B**). We manually classified each *k*-means group according to cell types displaying average HMR methylation ≤50% for each group. For example, in the "Hematopoietic" HMR group, blood cells uniquely display low methylation levels, whereas the "Early Developmental" HMR group is dominated by H1 ESCs. Likewise, a group of HMRs is specific to the "Fetal" developmental state compared to stem and adult cells. Using this analysis, we achieve remarkable resolution to distinguish HMRs that are unique between highly related cell types such as macrophage and neutrophil cells; further, we identify a more specific group of exclusive T and B cell HMRs.

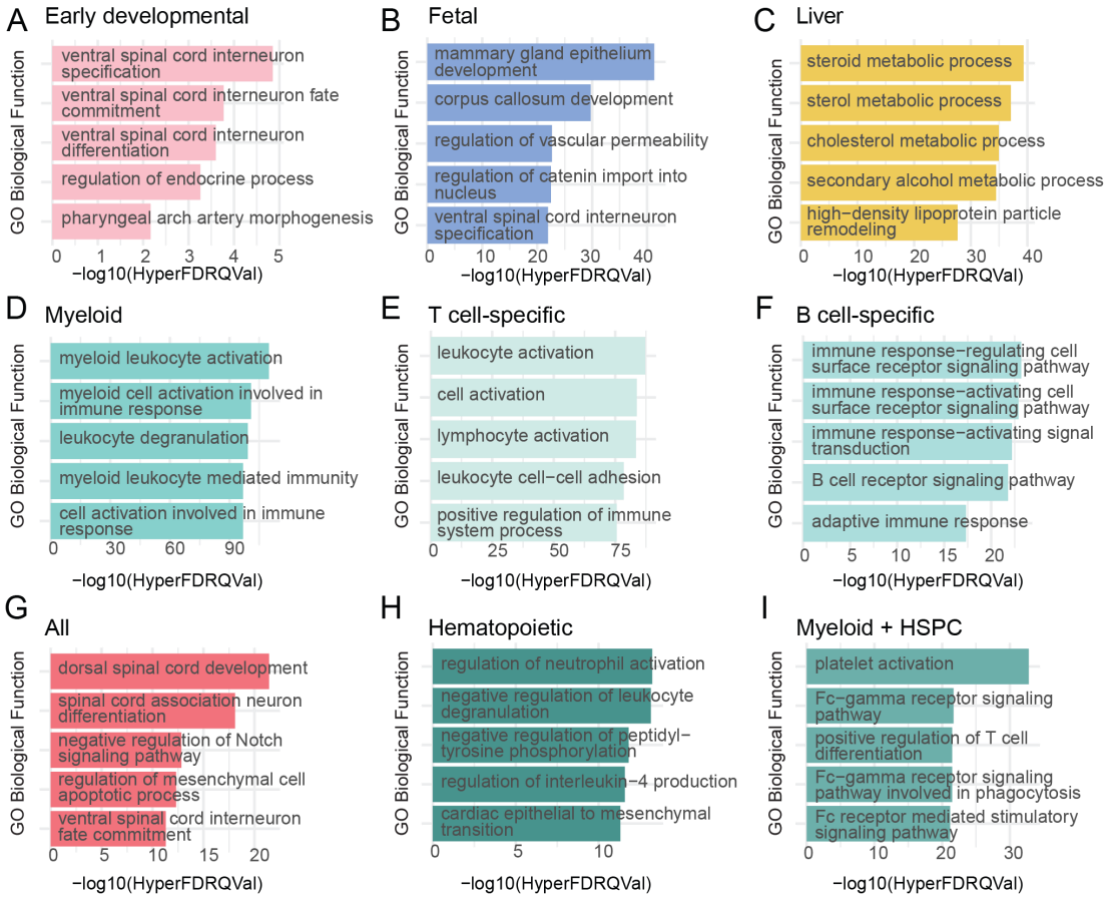
Since transcription factors (TFs) govern the functional progression and specialization of cell types, we performed TF motif enrichment analysis to understand the gene regulatory significance of each *k*-means group. Top motifs stratify strongly by *k*-means group (**Fig 3C**). Furthermore, representative

TFs from top results show *k*-means group-specific enrichment of canonical TFs indicative of their respective cell types. For example, the ubiquitous CTCF is enriched in the *All* group (110); pluripotency factors OCT4-Sox2-Nanog are primarily in the *Early Developmental k*-means group (111); CEBP, a factor important for myeloid development, is enriched exclusively in macrophages and neutrophils (myeloid cells) (112); and early B cell factor EBF2 is in the highly specific *B cell* group (113). Similarly, the retinoic acid receptor alpha (RARa) motif is highly enriched exclusively in the liver/adrenal-specific *Endocrine* group. The enrichment of cell specific transcription factors in cell specific HMRs confirmed expectations of our HMR group annotation strategy and highlights the ability to observe shared HMR patterns that reflect not only subsets of cell types but also developmental periods.

Given the specificity of the TF enrichment analysis supporting cell- and lineage-specific functions, we considered whether associated genes displayed similar biological specificity. We used GREAT ontology enrichment analysis to analyze sets of genes neighboring HMR groups defined by *k*-means clusters shown in **Fig 3B** (114). We show enrichment of distinct biological processes representative of the cell type and developmental stage associated with HMR groups (**Fig 7**). Interestingly, the least differentiated HMR group of *Early development* enriched for early morphogenic specification ontologies, while the intermediate HMR group defined by sharing between the blood cell types and stem and progenitor cells enriches for blood-related signaling ontologies. Additionally, the myeloid-specific enrichments show myeloid lineage specificity, whereas B cell-specific ontologies are enriched in the *B cell* group; this highlights the ability of HMRs to distinguish not only disparate lineages and developmental stages, but also highly related cell types. Together these data show that HMRs alone can recapitulate functional relationships between cell types. Furthermore, by comparing HMRs within and across lineages, we discovered that levels of HMR specificity can reflect deep developmental roots of gene regulation, capturing time point-specific branchpoints of development (**Fig 3 and 5**). For example, the hematopoietic *k*-means cluster contains a group of HMRs that are

shared between stem and progenitor cells as well as derived cell types (B cell, T cell, neutrophil, and macrophage), but not others. This data suggests that HMRs established at specific, early developmental timepoints are maintained in subsequent cellular states. We explore this in further detail below.

Figure 7. Bargraph of GREAT gene ontology results by methylation heatmap *k*-means cluster.
GREAT gene ontology enrichments are shown for cluster groups from the heatmap in Figure 3B (114). Results from the top 3 by hypergeometric FDR *q*-values are displayed. The *x*-axis shows the hypergeometric *q*-values. The cluster groups shown include (A) "Early developmental," (B) "Fetal," (C) "Liver," (D) "Myeloid," (E) "T cell-specific," (F) "B cell-specific," (G) "All," (H) "Hematopoietic," and (I) "Myeloid + HSPC."



HMRs accumulate and persist through subsequent developmental transitions.

Terminally differentiated cells exhibit between ~2-3.5 times the number of non-coding HMRs compared to embryonic stem cells (**Fig 3D**). While a minor subset of H1 ESC HMRs are cell type-

specific, most H1 ESC HMRs are highly shared across the cell types analyzed (**Fig 3B,** 14,223 merged HMRs that are shared among "All" cell types; of 18,235 H1 ESC non-coding HMRs, 2,616 are cell specific while 15,619 are shared with at least one other cell type, a 5.97-fold difference). Our comparative analysis further reveals specific HMR groups defined by developmental stage (fetal vs. adult, differentiated vs. undifferentiated), lineage, and cell type (**Fig 3B**). These data suggest a model whereby H1 ESCs supply a base HMR set to which additional HMRs are added at distinct lineage commitments through cell development. This is important because it suggests that a developmental hierarchy exists among HMRs and that HMRs accumulate as cells differentiate.

To determine whether progressive HMR establishment can be traced in developmentally derived cell types, we used pluripotent H1 ESCs, multipotent HSPCs, and terminally differentiated myeloid (macrophages) and lymphoid (B cells) lineage cells to construct a pseudo-time course (**Fig 3D**). In general, we observe that non-coding HMRs increase in number with increasing cell maturity. An increase of total HMRs could be explained by 1) a simple accumulation of additional HMRs, or 2) a net increase with high turnover of HMRs. To differentiate between these two modes of HMR expansion, we measured HMR overlap between either embryonic stem cells or hematopoietic stem cells and mature hematopoietic cell types. Of 18,235 HMRs observed in H1 ESCs, 11,959 (65.58%) were represented by HMRs in the total multipotent HSPC dataset. Of these 11,959 HMRs that were observed in both H1 ESCs and HSPCs, 11,310 (62.02%) and 10,285 (56.40%) were represented by HMRs in the macrophage and B cell datasets, respectively (**Fig 3E**). Next, of 34,605 HMRs established in HSPCs but absent in H1 ESCs, 27,312 (78.93%) and 15,185 (44.23%) were represented by HMRs in the macrophage and B cell datasets, respectively (**Fig 3E**).

These data show that a majority of the HMRs observed in differentiated cells (~60%) are established at early developmental stages and suggest a pattern of HMR accumulation in relation to developmental progression. In addition to acquiring new HMRs, macrophages retain a majority of

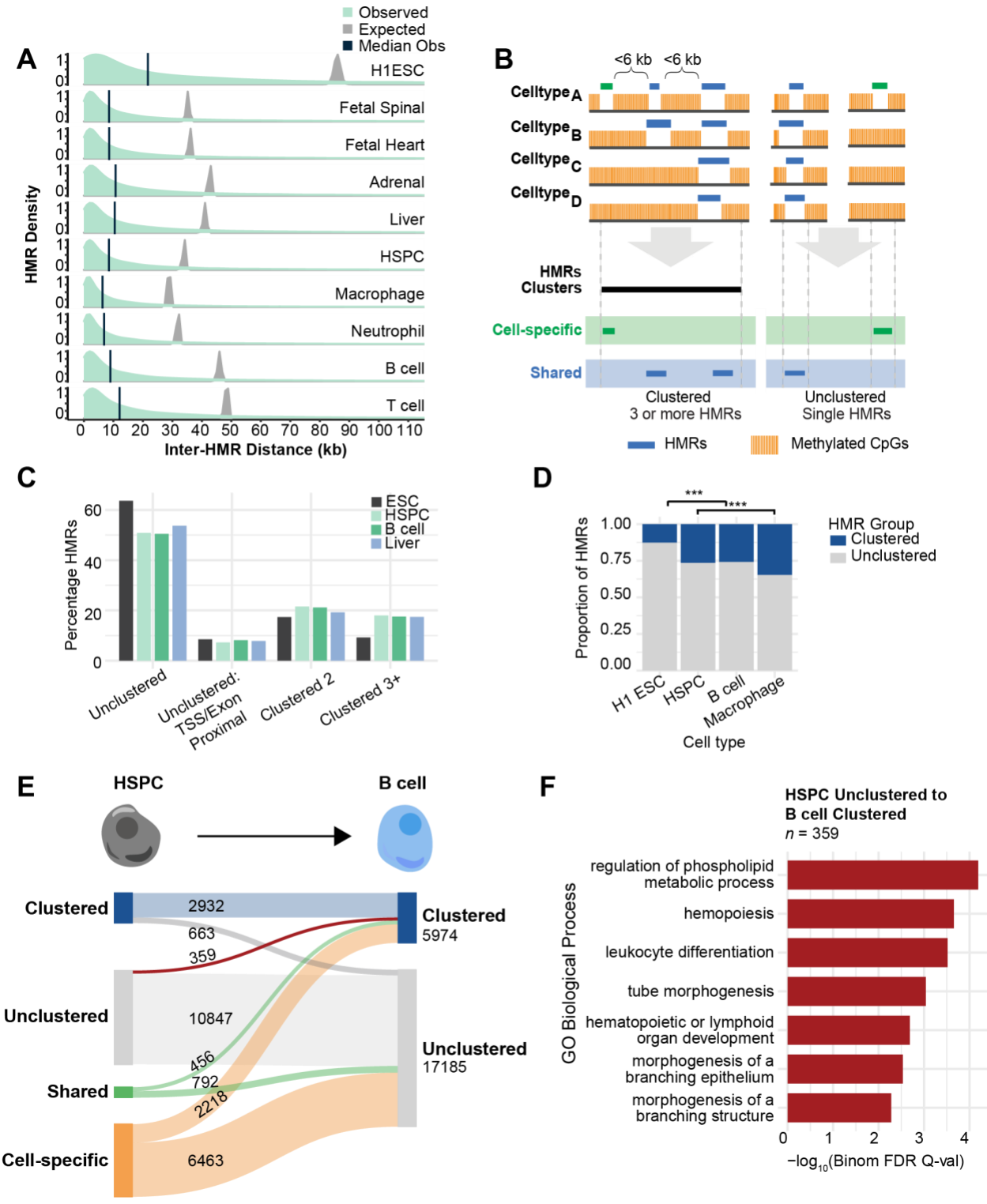
HMRs established in HSPCs, whereas B cells retain half as many HSPC-derived HMRs and fewer total HMRs compared to macrophages. This observation is consistent with previous studies demonstrating that lymphoid commitment and myeloid restriction requires re-methylation of specific early hematopoietic regulatory elements in parallel to demethylation of lymphoid-specific elements (115-117). Failure to remethylate these regions can result in a lineage priming imbalance favoring myeloid differentiation; thus, fewer HMRs are retained from HSPCs in B cells compared to macrophages. Despite this B cell remethylation of a subset of HSPC HMRs, we observe a general increase in HMRs across the hematopoietic lineage that supports a model where new HMRs are progressively established through successive developmental stages and persist through later stages of cell differentiation.

HMRs are non-randomly established into spatially organized clusters.

Locus-specific analysis of individual WGBS datasets indicates that multiple distinct HMRs are frequently located near one another, rather than being randomly distributed across linear genomic space (**Fig 3A**). Moreover, these HMR groups appear to be spatially organized with HMRs that are present in varying degrees of cell types and tissues, from developmentally constitutive to B cell specific. The example locus shown in **Fig 3A** depicts a group of adjacent HMRs near the *CD27* gene. *CD27* and several other genes surrounding the locus play a key role in B cell function (118-120). To quantify this HMR grouping phenomenon genome-wide, we calculated observed and expected distributions of inter-HMR distances utilizing the cell types represented in **Fig 3**. Expected distributions were simulated by random shuffling (n=10,000) HMRs across the genome for each dataset, excluding a blacklist of protein coding RefSeq TSSs (-2000/+1000 bp) and exons. HMR distances are significantly closer to each other than expected by random chance (**Fig 8A,** Wilcoxon rank sum, *p*-value < $2.2e^{-16}$). Interestingly, differentiated cell types consistently show lower expected and observed distances (~40-50 kb and ~12 kb, respectively) compared to those of H1 ESCs, which

feature the largest inter-HMR distances; this is consistent with having fewer HMRs overall, supporting its role as a basal HMR set.

Figure 8. HMRs cluster more than expected.
A) Distribution plots of inter-HMR distances by cell type. The green distributions represent observed values from HMR datasets per cell type. Vertical navy bars show median values. Grey distributions show expected values by random shuffling across the non-coding genome. For each cell type, the expected and observed distributions were determined to be significantly different by the Wilcoxon rank sum test. All *p*-values were reported as zero ($p < 2e\text{-}16$) with a range of X^2 values from 1.4337×10^8-4.4508×10^8. (B) Diagram of HMR clustering and cell specificity workflow. HMRs are annotated for clustering behavior and/or cell specificity. Non-coding HMR datasets are defined by HMRs that do not overlap RefSeq protein-coding TSSs (TSS -2000/+1000) and exons. Clustering refers to groups of HMRs in a cell type that are located a maximum of 6 kb end-to-end from the next HMR, linking 3 or more HMRs; clusters cannot cross TSSs or exons. Unclustered HMRs are defined as non-coding HMRs that are not within 6 kb of any other non-coding or TSS/exon-overlapping HMR. Cell specificity is also defined, with any base pair overlap between HMRs constituting overlap. (C) Bar graph of HMR clustering annotations discussed in (B) and **Fig S6** as percentages of total HMRs by cell type. Selected cell types represent members of the hematopoietic and hepatic lineages. Colors reflect cell types representing different developmental stages and lineages. (D) Bar graph of proportion of cell type HMRs that are clustered HMRs (3+ HMRs) vs unclustered. Total values are calculated as [#unclustered + #clustered]. (E) Sankey diagram showing the flow of B cell HMRs. B cell HMRs are divided on the right of the panel into clustering groups. The left shows HSPC HMRs that overlap B cell HMRs, and are hierarchically categorized as *clustered HSPC* HMR, *unclustered HSPC HMR*, *shared*, or *cell specific*. To define cell specificity, B cell HMRs were compared to datasets from adrenal gland, H1 ESC, HSPC, fetal spinal, fetal heart, liver, macrophage, neutrophil, and T cell. (F) The bar graph shows the top biological process gene ontology results for the Sankey group of HMRs that progress from *HSPC unclustered* to *B cell clustered* (indicated in red). Results from GREAT Gene Ontology using default background and gene assignment settings are represented by bars showing binomial *q*-value (114).



These data suggest that clustered HMRs play a distinct regulatory role compared to their unclustered counterparts. To characterize the features that distinguish “clustered” and “unclustered” HMRs we

first determined a set of heuristic criteria to define clusters (**Fig 8B**). We plotted per-cell type distributions for non-coding inter-HMR distances and measured distance quantiles. From this, we analyzed "end-to-end" cluster lengths based on three maximum linking values: the ≤12.5 kb stitching distance commonly used in ChIP-seq-based super-enhancer studies (71, 104, 121-123); the approximate mean inter-HMR distance of 11 kb and ≤6 kb which represents the median inter-HMR distance after filtering for values under 50 kb (**Table 3**). Previous studies that have characterized clustered super-enhancers have used a common linking distance threshold of 12.5 kb. This distance was reportedly selected for its ability to qualitatively link high signal regions together while avoiding inclusion of lower signal peaks. Thus, the super-enhancer definition is reliant upon signal intensity and distribution. However, HMRs are defined by a bimodal methylation signal distribution, and such a distance applied to methylation data results in extraneously long stitched regions, the biological function of which is difficult to assign; some exceed 1Mb, which can result from HMRs spread across gene deserts, or large topological domains with low methylation levels or CpG frequency. By comparison, a linking distance of 6 kb results in stitched regions with an overall mean length of ~10 kb, which is consistent with other clustered enhancer annotations such as stretch and super-enhancers (**Table 4, Fig 9)** (73, 104). Using a linking distance of 6 kb, we determined the fraction of HMRs that are clustered or unclustered for a subset of cell types, including H1 ESC, HSPC, B cell and Liver (**Fig 8C**). To avoid confounding contributions of promoter characteristics to our analysis, clusters were not allowed to cross TSSs or exons. At a 6 kb threshold, non-TSS/exon HMR groupings that exist as pairs or as clusters of 3 or more constitute ~35% of all HMRs. For the rest of this paper, “clusters” refer to clusters with 3 or more HMRs (see annotation strategy in **Fig 10**).

Figure 9. HMR cluster lengths are consistent across cell types.
The graph shows the lengths of HMR clusters, end-to-end, per cell type. Data is represented by both a violin plot and boxplot. The boxplot shows the interquartile range, and the bold black line shows the median value per cell type. The red dotted line shows the value 10,000 bp, which approximates the mean cluster length of 9764.59 bp, measured across the cell types: H ESC, fetal heart, fetal spinal cord, adrenal gland, liver, HSPC, macrophage, neutrophil, T cell, and B cell.

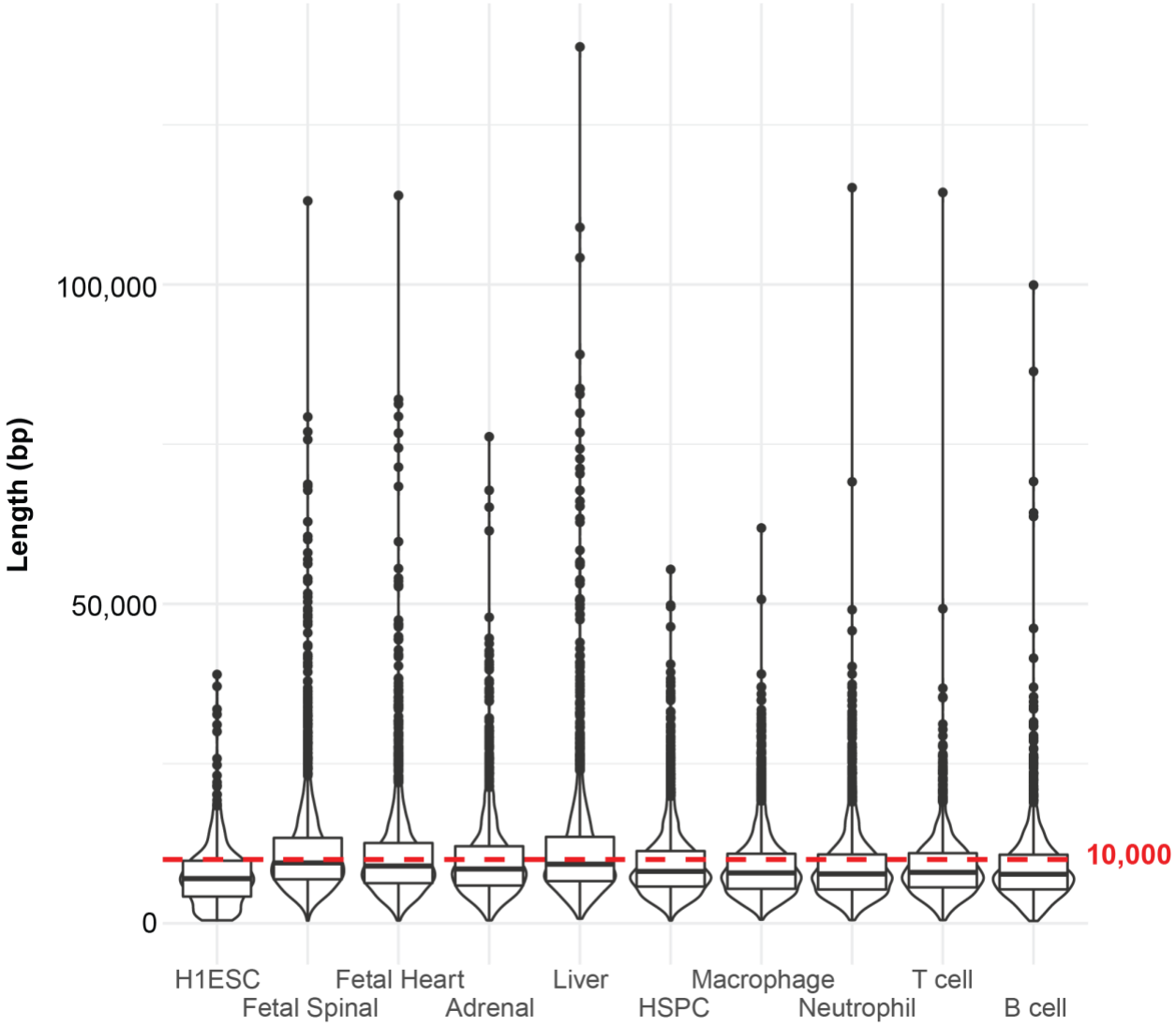


Table 3. Inter-HMR lengths by cell type.

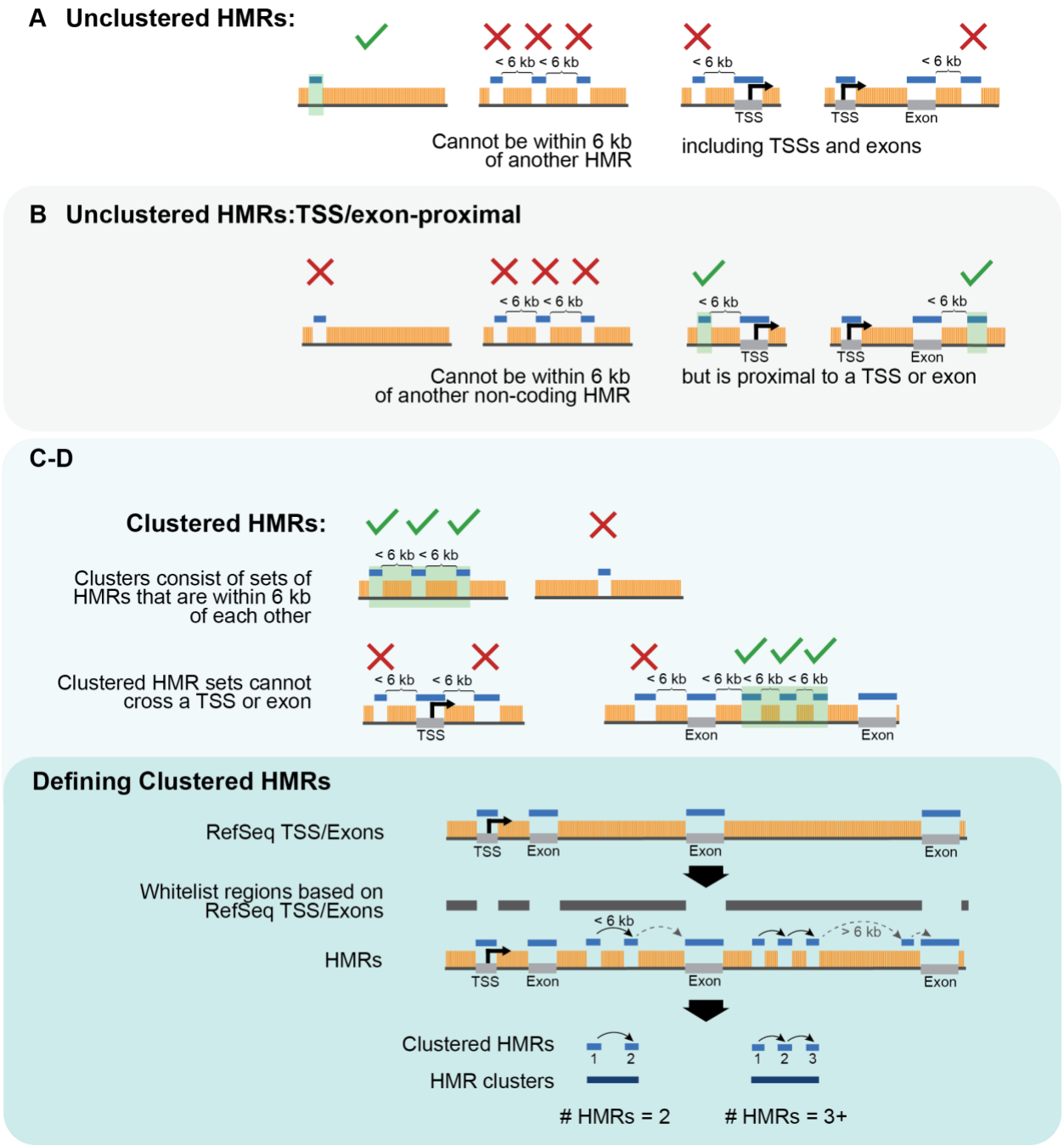
Quantile	H1 ESC	Fetal Spinal	Fetal Heart	Adrenal	Liver

0.1	227	881	761	673	633
0.2	1283	1770	1626	1678	1508
0.3	2953.8	2960	2787	2969	2716.8
0.4	5626.6	4574	4329	4777	4441.4
0.45	7102.65	5485	5309	5990	5540
0.5	8921	6555	6361	7311	6846
0.55	11013.15	7805	7689	8799.3	8461
0.6	13299.8	9237	9183	10524	10148.2
0.7	18970	13267	13231	15105.6	14803
0.8	26779	19213	19414	21739.2	21582
0.9	37041.2	29618	29903	32016.2	31880.6
Quantile	HSPC	Macrophage	Neutrophil	T cell	B cell
0.1	652	535.9	523	715	436.3
0.2	1469	1152.8	1190	1789	1259
0.3	2618	2025.7	2123	3329.4	2410
0.4	4222.4	3284	3472	5214	4030
0.45	5205	4078.55	4298	6397.3	4998
0.5	6330	5001	5265	7728	6142
0.55	7668.15	6080.45	6409.95	9361	7474
0.6	9208	7375	7824.4	11244.8	9112.8
0.7	13361	10919	11583	15820.8	13573.1
0.8	19384	16430.2	17323	22468.6	19841.2
0.9	29702	26317.3	27366	32882.2	30592.7

Table 4. Clustering group region counts by clustering distance (bp).

Linking Distance	Unclustered	Clustered
6,000	17,185	5,974
11,000	12,441	9,044
12,500	11,668	9,717

Figure 10. Schematic of HMR definitions and annotation.
Visual graphic of HMR definitions for groups: (A) unclustered, (B) unclustered: TSS/exon proximal, (C) clusters of 2 HMRs, and (D) clusters of 3+ HMRs. Gene tracks are not to scale.



As demonstrated in **Fig 3A**, a typical cluster consists of multiple HMRs with different levels of cell type specificity between them—broadly shared (developmentally constitutive), lineage-shared or cell-specific. This means that a cluster identified in one cell type may not exist across all cell types. As the formation of clusters is contingent on the addition of new HMRs near existing HMRs, most HMR clusters (~35-40%) contain at least one lineage- and/or cell type-specific HMR. Given that HMRs

accumulate over developmental timelines, this observation raises the possibility that, as cells differentiate, HMRs are preferentially added to clusters in a lineage-specific manner. Indeed, we observe a positive correlation between clustering and developmental state. Clustering percentage increases as development progresses (**Fig 8D**, *H1 ESC to HSPC* & *HSPC to Macrophage*: $p<2.2x10^{-16}$), and this is accompanied by a relative decrease in unclustered HMRs. Tracking these HMRs temporally for each pseudo-timepoint reveals that a substantial fraction of early-established HMRs is joined by additional HMRs in subsequent developmental states. The establishment of new HMRs near existing HMRs can lead to clustering, where an HMR may be classified as unclustered at an early timepoint but become clustered in a differentiated cell type (**Fig 8E**). These growing clusters of HMRs are often in proximity to lineage-specific genes, as suggested by gene ontology analysis (**Fig 8F**). Altogether, these data show that HMRs can be broadly distinguished by 1) the number of cell types that share them—a corollary of temporal establishment or developmental time—and 2) their clustering behavior, which may reflect a collective and unique developmental function that distinguishes clustered HMRs from other types of genomic regions.

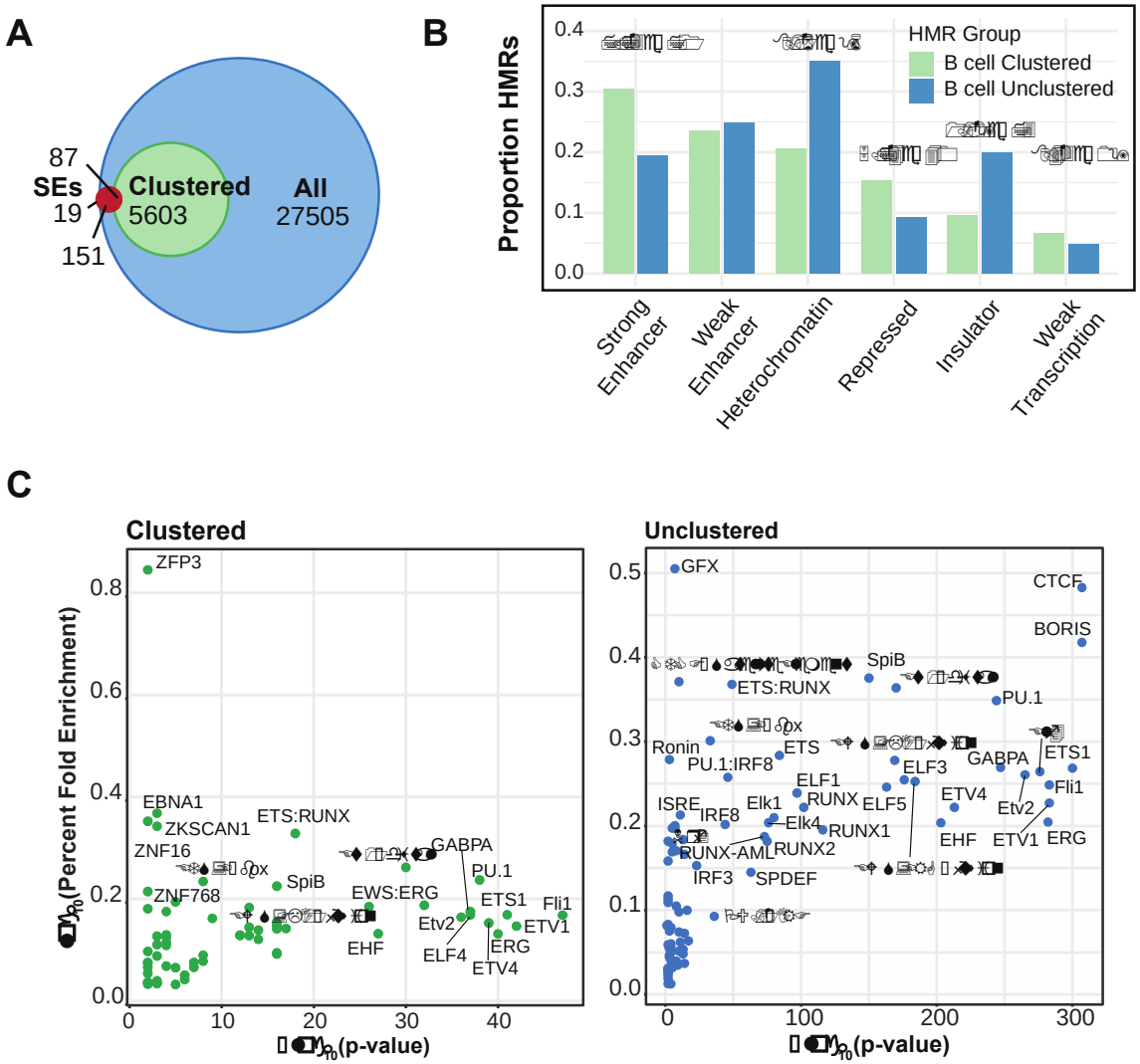
Clustered HMRs are functionally distinct from unclustered HMRs.

Sequencing approaches have enabled the discovery of many spatially clustered regulatory elements genome-wide using chromatin accessibility (73), histone modifications (124-129), and transcription factor binding (71, 104). More recently, clustering of enhancers has been commonly associated with concepts such as super-enhancers (SEs) (104), stretch enhancers (73), shadow enhancers (130, 131) and locus control regions (LCR) (132, 133), which are thought to provide regulatory additivity, synergy, and redundancy to their target genes in a tissue-specific manner.

Comparison of clustered B cell HMRs with histone H3K27ac-defined B cell super-enhancers shows that, while the majority of super-enhancers coincide with HMRs (both clustered and unclustered), only 1.5% of clustered HMRs overlap super-enhancers (**Fig 3A**) (104). This discrepancy may be

explained by the observation that only a fraction of SEs exists as clusters in linear genomic space. Indeed, 15% of SEs from Whyte *et al*. are singletons and only 196 of 1,355 stitched murine ESC enhancers are SEs (71, 72). Thus, super-enhancers do not exclusively consist of clustered enhancers, and ChIP-seq defined enhancer clusters are not exclusively SEs. Clustered HMRs are more frequent than SEs, and their existence raises the question of whether they represent distinct functional characteristics compared to their unclustered HMR counterparts.

Figure 11. Clustered HMRs show distinct enhancer-associated characteristics compared to unclustered HMRs.
(A) Venn diagram showing partially overlapping sets between three region datasets: All B cell HMRs (blue line); clustered B cell HMRs (green line; subset of All); and GM12878 super-enhancers (solid red circle) (71). GM12878 is a tier 1 ENCODE lymphoblastoid cell line derived from EBV immortalized B cells. (B) Bar graph of HMR overlap with selected ChromHMM annotations: *strong enhancer*, *weak enhancer*, *heterochromatin*, *repressed*, *insulator*, *weak transcription*. The height of the bars represents the fraction of clustered and unclustered HMRs that overlap each annotation. Z-test of proportion *p*-values are shown, comparing HMR group proportion values for each ChromHMM annotation. (C) TF motif enrichment in clustered (left; green) and unclustered (right; blue) HMRs. Results are plotted as $-\log_{10}p$-value by fold enrichment, measured as percentage of target regions containing motif divided by the percentage of background regions. Background represents all clustered and unclustered HMRs.



To address this question, we used ChromHMM annotations to functionally categorize HMRs based on clustering behavior in B cells (**Fig 11B**) (124, 125). Notably, clustered HMRs are enriched for "strong enhancers" (X^2=316.66, p=7.725x10^{-71}) while unclustered HMRs show higher enrichment of "heterochromatin" (X^2=432.37, p=8.159x10^{-96}) and "insulators" (X^2=329.57, p=1.191x10^{-73}). This suggests clustered HMRs are enriched for active regulatory regions while unclustered HMRs tag elements involved in three-dimensional chromatin structure. This result is corroborated by the

strong enrichment of the CTCF motif in unclustered HMRs, while both clustered and unclustered B cell HMRs show comparable enrichment of lymphoid-relevant transcription factors, including PU.1, SpiB, and ETS family members (**Fig 11C**).

Given the enrichment of strong enhancer annotations in clustered HMRs, we investigated their transcriptional regulatory activity by comparing with our recently published ATAC-STARR-seq data for immortalized B cells (**Fig 12A**) (134). ATAC-STARR-seq is a massively parallel reporter assay that uses Tn5 transposase to selectively clone accessible DNA from native chromatin into a plasmid-based reporter to test accessible chromatin regions for active and silent regulatory activity (134, 135). Since a majority of B cell HMRs overlap accessible chromatin regions in lymphoblastoid cells (**Fig 13**), we measured the proportion of HMRs that contain an activator or silencer (**Fig 12A**). Despite being fewer in number, clustered HMRs contain a significantly higher proportion of transcriptional regulators, including both activators and silencers (p = 2.39x10^{-13} and 0.0106, respectively), than unclustered HMRs.

Figure 12. Clustered HMRs are enriched for active regulatory elements compared to unclustered HMRs.
(A) Boxplot of ATAC-STARR-seq regulatory element overlap by clustered and unclustered HMRs. Overlap is measured at the unit of HMRs, and values depict fraction of total HMRs that contain a regulatory element. A Wilcoxon rank sum test was used to determine statistical significance. (B) Point and line graph of the proportion of HMRs near an expressed gene at different TSS distances. HMRs are grouped by HMR clusters that contain a cell specific HMR and unclustered cell specific HMRs. Denominators for the HMR clusters and unclustered HMR groups are 444 and 1621, respectively. Counts below the graph represent the cumulative amount of genes below each threshold per HMR group. *p*-values are derived from a z-test of proportions to test the fraction of HMRs represented by HMR-single nearest neighbor gene pairs below each threshold distance. (C) Boxplot of TPM values (derived from GM12878 cell line data) of nearest neighbor RefSeq protein-coding genes to clustered and unclustered HMRs. Two nearest neighbor genes (with TPM > 0) per HMR were filtered for TAD boundary crossing. Statistical significance was measured by a Wilcoxon rank sum test in R using the *wilcox.test() function*. (D) Multiple alignment of region around the *CD27* locus showing methylation and HMR tracks across 6 cell types: H1 ESC, hematopoietic stem and progenitor cells, macrophage, neutrophil, B cell, and T cell. Methylation tracks are represented by orange vertical bars showing methylation value per CpG site. HMRs are shown by dark blue horizontal bars. Below the multiple alignment, Hi-C interaction score data is represented by heatmap triangles representing interaction matrices for GM12878s and H9 ESC cells. Values for the Hi-C data are derived from .hic interaction matrix files. The *plotgardener* R package was used to generate the genome browser snapshot (106).

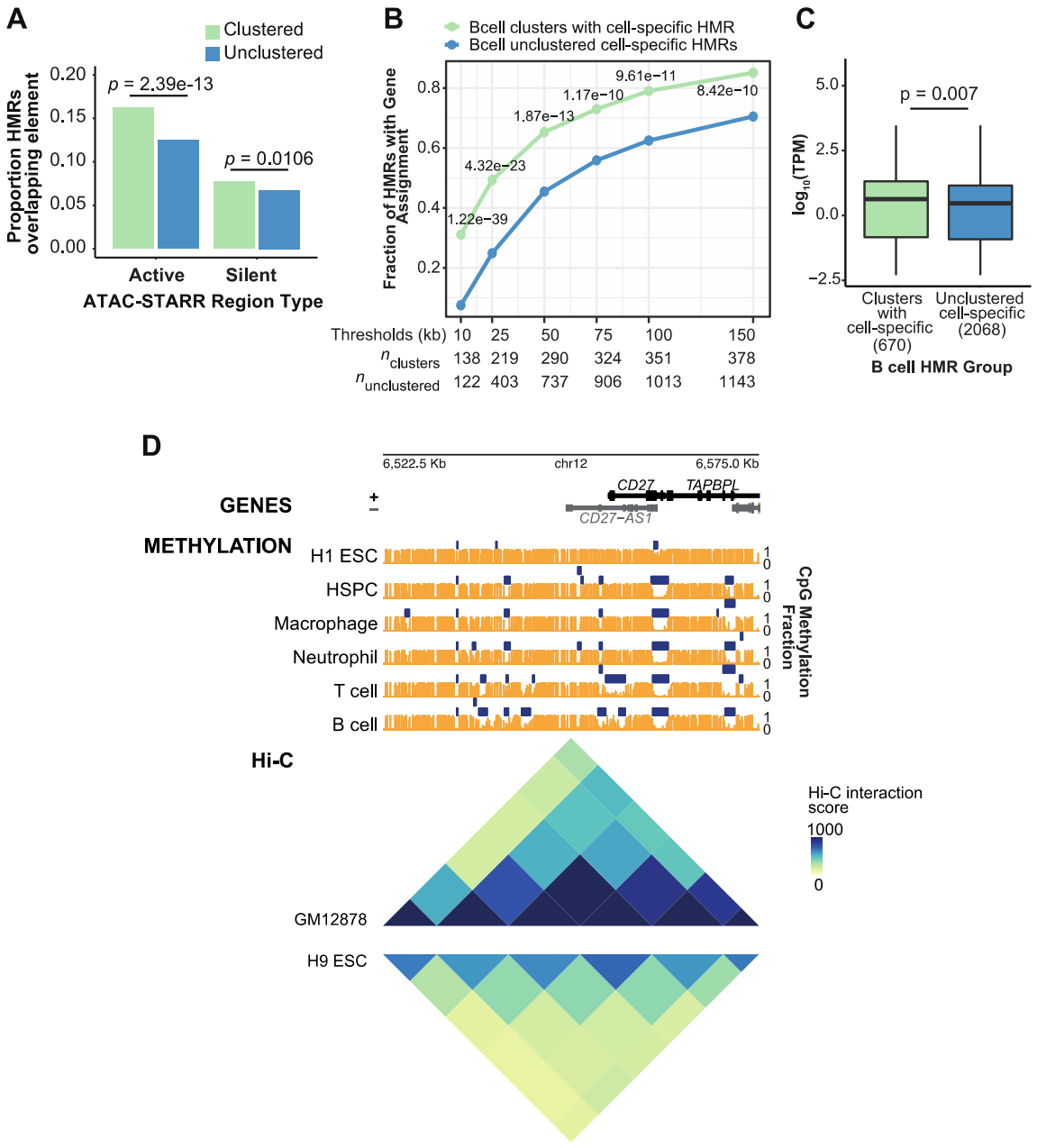
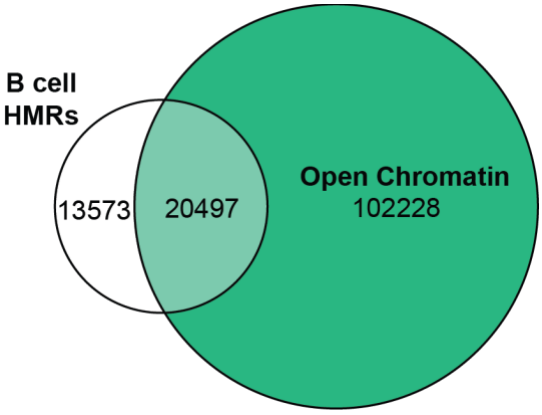

A
Clustered
Unclustered
Proportion HMRs overlapping element
0.20
0.15
0.10
0.05
0.00
p = 2.39e-13
p = 0.0106
Active
Silent
ATAC-STARR Region Type
B
Bcell clusters with cell-specific HMR
Bcell unclustered cell-specific HMRs
Fraction of HMRs with Gene Assignment
0.8
0.6
0.4
0.2
9.61e−11
1.17e−10
8.42e−10
1.87e−13
4.32e−23
1.22e−39
Thresholds (kb) 10 25 50 75 100 150
n clusters 138 219 290 324 351 378
n unclustered 122 403 737 906 1013 1143
C
p = 0.007
log10(TPM)
5.0
2.5
0.0
−2.5
Clusters with cell-specific (670)
Unclustered cell-specific (2068)
B cell HMR Group
D
6,522.5 Kb
chr12
6,575.0 Kb
GENES
CD27
TAPBPL
CD27−AS1
METHYLATION
H1 ESC
HSPC
Macrophage
Neutrophil
T cell
B cell
CpG Methylation Fraction
Hi-C
Hi-C interaction score
1000
0
GM12878
H9 ESC

Figure 13. Euler plot comparing B cell HMRs with open chromatin.
Euler plot of all B cell HMRs and open chromatin defined by DNase I hypersensitivity sites in GM12878 cells. The DNase file was downloaded from the UCSC Genome Browser Table Browser using the following main settings: clade: "mammal"; genome: "human"; assembly: "Feb 2009 (GRCh37/hg19)"; group: "Regulation"; track: "Duke DNaseI HS"; table: GM12878 Pk (wgEncodeOpenChromDnaseGm12878Pk)" [ENCODE file ID: ENCFF001UVC]. The values, 13573 and 20497, represent count values for HMRs. The value 102228 represetns a count for open chromatin regions.



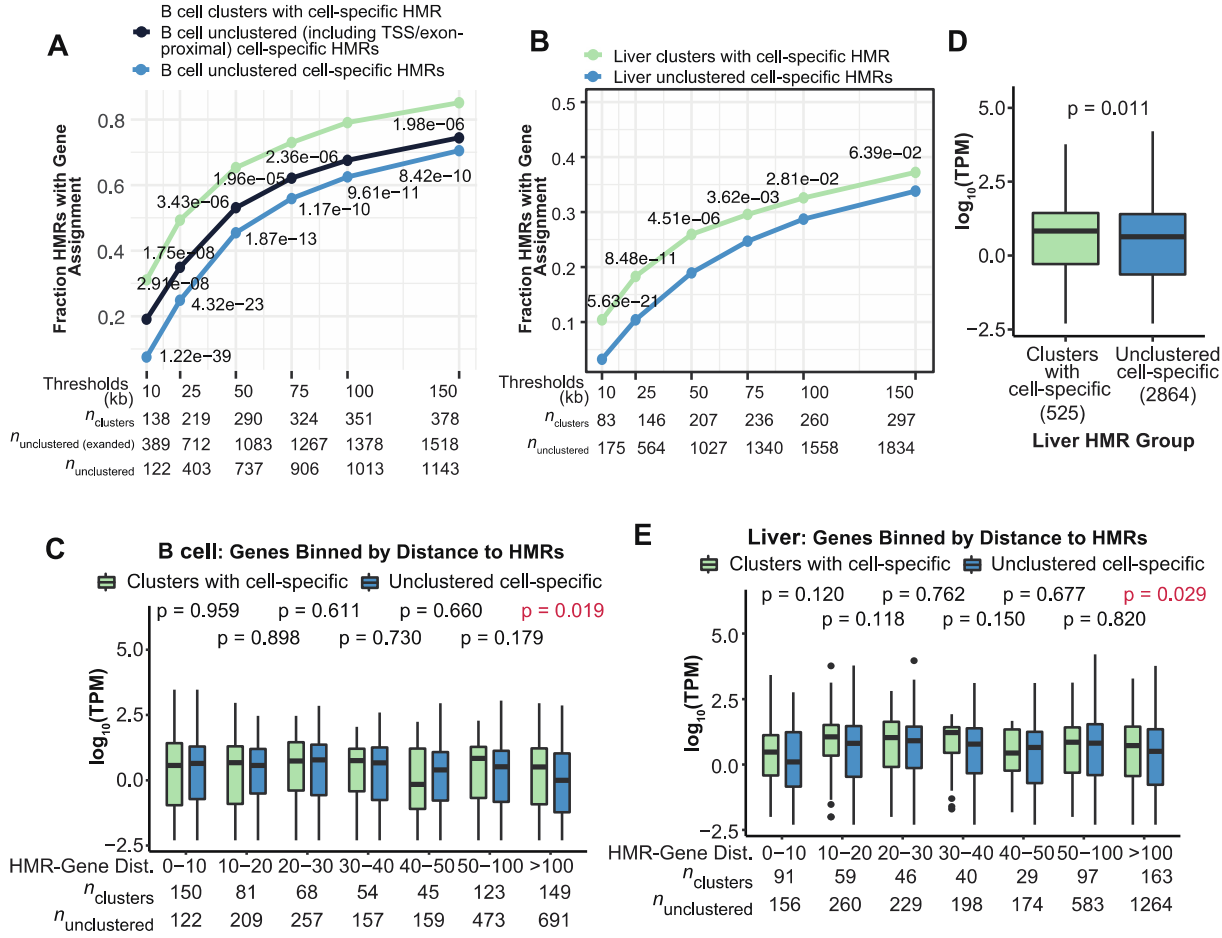
Based on the finding that clustered HMRs are enriched for both strong enhancer annotations and "activators" defined by ATAC-STARR-seq (**Fig 12A**), we hypothesized that clustered HMRs are more likely to be associated with active genes compared to unclustered HMRs. To address this question, we defined pairs of HMRs and their nearest neighbor genes, measuring the proportion of cell-specific clustered and unclustered HMRs that tag the nearest "active" (TPM>0) gene at different threshold HMR-TSS distances (**Fig 12B**). To pair HMRs with their nearest neighbor active gene, we used a gene assignment strategy that identifies the nearest expressed neighboring gene within a topologically associated domain (TAD) containing both the gene and the HMR(s). A recent study showed that a combination of nearest neighbor assignment in conjunction with a minimum expression threshold increased associated-gene prediction accuracy above several gene assignment methods, including the commonly utilized simple nearest neighbor (87). Using this assignment strategy, we paired HMR groups with lymphoblastoid RNA-seq (GM12878) data from ENCODE as a proxy for B cells (136). Focusing on B cell specific HMRs, we observed a significantly higher proportion of clustered HMRs near active genes compared to unclustered HMRs at all observed distance thresholds (**Fig 12B;** all *p*-

values <8.42e-10). While our unclustered definition omits HMRs that are within 6 kb of TSSs or exons, these observations remain consistent when TSS/Exon proximal HMRs are included (**Fig 14A**; all *p*-values <1.96e-5). Similar results were obtained for the same analysis performed on liver HMRs (**Fig 14B**).

We further reasoned that target genes of clustered HMRs display increased transcriptional output compared to those of unclustered HMRs. To define putative HMR target genes, we used a similar approach to the gene assignment strategy discussed above that incorporates both a TAD and expression filter (TPM > 0); due to uncertainty that the nearest gene is a true positive, we considered two nearest neighbors for gene assignment. Using this approach, we observe that genes assigned to clustered B cell HMRs show a significantly higher distribution of transcript levels compared to those near unclustered HMRs ($p = 0.007$; **Fig 12C**) (136). However, when these comparisons are binned by HMR-gene distance, we do not observe significant differences in gene expression across bins, except in the most distal (≥100 kb) HMR-gene distance bin (**Fig 14C**). We performed the same analysis for liver clustered HMR target genes, finding the pattern is consistent across cell types (**Fig 14D-E**). These observations suggest that the functional distinction of clustered HMRs compared to unclustered HMRs is a general phenomenon.

Figure 14. HMR proportions near active genes and boxplots comparing gene expression and distance near clustered and unclustered HMRs.
(A) Point and line graph of the percentage of HMRs that are found in HMR-gene single nearest neighbor pairs at different distances. HMRs are grouped by HMR clusters that contain a cell-specific HMR and unclustered cell-specific HMRs. Denominators for the HMR clusters, unclustered (including TSS/exon-proximal), and unclustered HMR groups are 444, 2040, and 1621, respectively. *p*-values are derived from a z-test of proportions to test the fraction of HMRs represented below each threshold distance. (B) Point and line graph of the percentage of HMRs that are found in HMR-gene single nearest neighbor pairs at different distances. HMRs are grouped by HMR clusters that contain a cell-specific HMR and unclustered cell-specific HMRs. Denominators for the HMR clusters and unclustered HMR groups are 798 and 5424, respectively. Counts below the graph represent the cumulative amount of genes below each threshold per HMR group. *P*-values are derived from a z-test of proportions to test the fraction of HMRs represented below each threshold distance. (C) Boxplot of normalized read counts of nearest neighbor RefSeq protein-coding genes to clustered and unclustered Liver HMRs. Nearest neighbor genes were filtered for TAD boundary crossing. Results

for Liver are also displayed in (D) for all genes, but binned by distance between the HMR and nearest gene. Statistical significance was measured by a Wilcoxon rank sum test.



Given the relationship between clustered HMRs and gene activity, we considered whether the appearance of clustered HMRs in differentiated cells accompanies changes in chromatin conformation. We used publicly available Hi-C data to compare long range chromatin contacts around the *CD27* locus between embryonic stem cells and differentiated B cells (**Fig 12D**). This locus provides a representative example of a cluster of HMRs that accumulates HMRs with increasing developmental specificity. Here, we observe that the accumulation of immune cell specific HMRs coincides with chromatin conformation changes as indicated by increased frequency of Hi-C

interactions (**Fig 12D**). As the region accumulates clustered HMRs through cell development, new chromatin contacts are created around the newly established HMRs (137, 138), indicating the functional importance of the spatial proximity of clustered HMRs. Altogether, these results argue that combinatorial HMR establishment and HMR history relates to chromatin conformation changes that accompany cell differentiation (see Discussion).

Non-coding HMR patterns are highly enriched for genetic variants linked to specific clinical phenotypes.

Genome-wide associations studies (GWAS) have demonstrated that a substantial portion of human phenotype-associated single nucleotide polymorphisms (SNPs) is located in functional regulatory elements (139-143). Integration of GWAS with functional genomic data reveals that disease risk variants also localize primarily within cell type-specific enhancers of disease-relevant tissues (84). Studies examining the relationships between disease loci and molecular phenotypes such as gene expression, chromatin accessibility or the DNAme status of *cis*-acting enhancers have identified a strong connection between non-coding genetic variants and epigenetic regulation (144-146). Based on these previous studies, we expected a SNP enrichment among HMR patterns that would associate with various traits. We therefore asked whether specific HMR patterns harbor genetic variants linked to distinct clinical phenotypes, and, in turn, whether these relationships can inform the functional significance of different HMR patterns.

We reasoned that GWAS SNPs could be leveraged to reveal genetic variants in HMRs of critical importance to normal cell development and function. As in **Fig 3D-E**, we defined B cell HMRs that are H1 ESC-derived (developmentally constitutive), HSPC-derived (lineage-shared) or B cell-specific. GWAS SNPs not only reflect trait-associated genetic variation, but also GWAS summary statistics can be used to estimate partitioned genetic heritability of traits assigned to subsets of the genome, based on the assumption that regions with higher quantities of SNPs in high linkage disequilibrium are

more likely to capture a causative variant. We used stratified LD score regression (S-LDSC) to perform partitioned heritability analysis from GWAS summary statistics of 79 traits and clinical lab values representing a range of organ systems ((89) **Table 5**). We found significant enrichment of trait heritability within lineage- and cell-specific HMRs (**Fig 15A**).

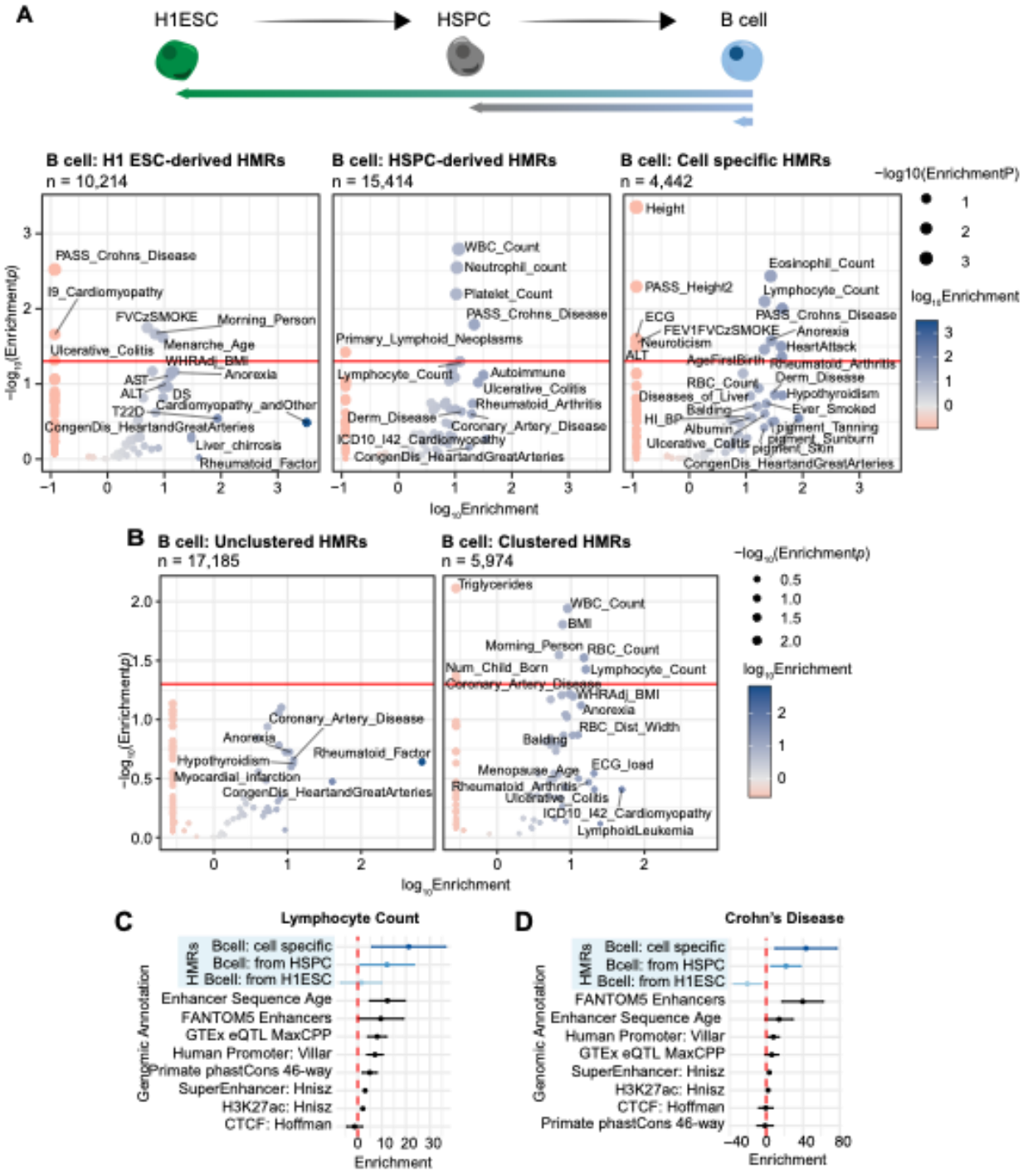
Table 5. List of 79 summary statistic files used for S-LDSC analyses.

Traits:
Albumin
ALP
ALT
Angina_byDoctor
Apolipoprotein_B
AST
blood_EOSINOPHIL_COUNT
blood_PLATELET_COUNT
blood_RBC_DISTRIB_WIDTH
blood_RED_COUNT
blood_WHITE_COUNT
bmd_HEEL_TSCOREz
body_BALDING1
body_BMIz
body_HEIGHTz
body_WHRadjBMIz
bp_SYSTOLICadjMEDz
Cadiomyopathy_andOther
CardiacArrythm
Cholesterol
Coffee_type
Congen_Heart_andGreatArteries
cov_EDU_YEARS
cov_SMOKING_STATUS
disease_AID_SURE
disease_ALLERGY_ECZEMA_DIAGNOSED
disease_DERMATOLOGY
disease_HI_CHOL_SELF_REP
disease_HYPOTHYROIDISM_SELF_REP
disease_RESPIRATORY_ENT

Diseases_of_liver
disease_T2D
ECG_load
ECG_phaseTime
ECG
Haematocrit_percentage
HeartAttack_byDoctor
HighBloodPressure_byDoctor
I25_chronicIHD
I9_Cardiomyopathy
I9_IHD_wideDefinition
ICD10_I42_Cardiomyopathy
ICD10_I48_atrialFibrillationAndFlutter
IGF1
Liver_chirrosis
lung_FEV1FVCzSMOKE
lung_FVCzSMOKE
Lymphocyte_count
LymphoidLeukemia
mental_NEUROTICISM
Myocardial_infarction
Neutrophil_count
other_MORNINGPERSON
PASS_AgeFirstBirth
PASS_Anorexia
PASS_Autism
PASS_BMI1
PASS_Coronary_Artery_Disease
PASS_Crohns_Disease
PASS_DS
PASS_Ever_Smoked
PASS_HDL
PASS_Height1
PASS_LDL
PASS_NumberChildrenEverBorn
PASS_Rheumatoid_Arthritis
PASS_Schizophrenia
PASS_Type_2_Diabetes
PASS_Ulcerative_Colitis
PASS_Years_of_Education2

pigment_HAIR
pigment_SKIN
pigment_SUNBURN
pigment_TANNING
Primary_lymphoid_neoplasms
repro_MENARCHE_AGE
repro_MENOPAUSE_AGE
RheumatoidFactor
Triglycerides

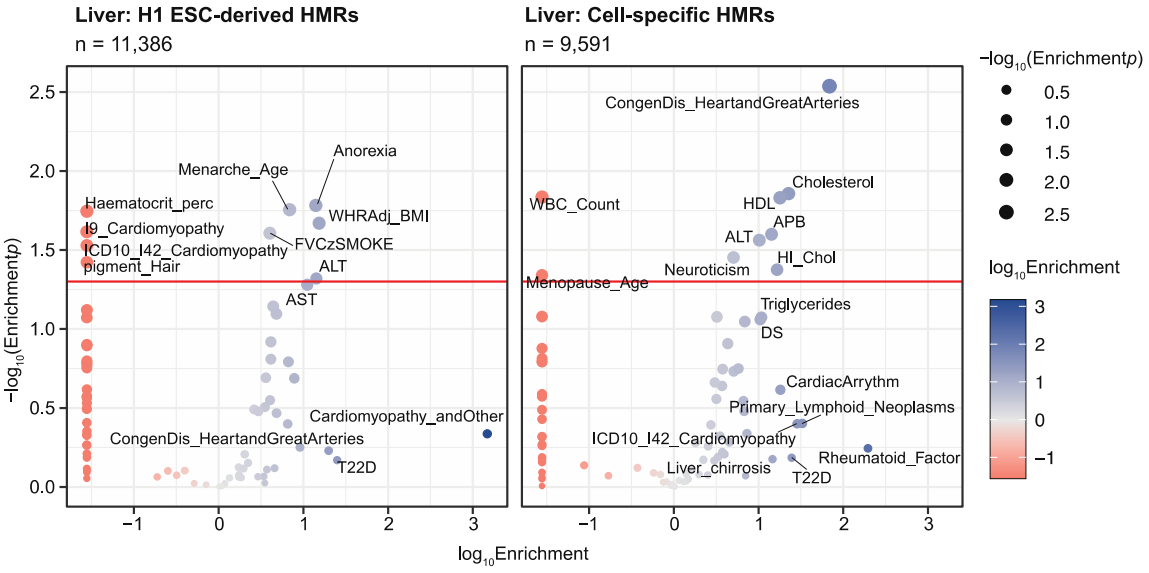
Figure 15. S-LDSC identifies HMR annotation-specific trait enrichments.
(A) Volcano-style plots of S-LDSC partitioned heritability results across 79 traits are shown for three *B cell* HMR groups*: H1 ESC-derived, HSPC-derived, and cell specific*. HMRs are ordered by the developmentally distinct cell type in which they were established. Each HMR group was tested for enrichment of genetic heritability with a standard set of 98 base annotations against traits that include both clinical diseases as well as clinical lab values. Negative enrichment values were clipped to the lowest positive enrichment value for each row of plots (A: 0.1174537; B: 0.25754925). The size of each point represents the $-\log_{10}p$-value of the enrichment, and the color shows the $\log_{10}$enrichment value. Points with a p-value <= 0.05 or an enrichment > 10 are labeled by their trait name where available. (B) Further partitioned heritability analysis applied to B cell HMRs grouped only by clustering behavior is also represented. (C) Point and line plot of S-LDSC enrichment values by annotation group for "Lymphocyte Count". These graphs include data from developmentally derived B cell HMRs compared against other enhancer-associated groups, including ancient human enhancer sequence age, FANTOM 5 enhancers, eQTLs, super-enhancers, and the H3K27ac histone mark. Genomic controls were also included, such as phastCons 46-way annotations as well as promoters and CTCF sites. The x-axis represents enrichment values, and the y-axis displays genomic annotations. Points show enrichment estimates and lines display 95% confidence intervals. The red line marks an enrichment score of 0. (D) Point and line plot of S-LDSC enrichment values by annotation group for "Crohn's Disease".



More specifically, we found that trait specificity not only stratifies by but also increases with HMR specificity. For example, H1 ESC-derived B cell HMRs are nominally enriched for traits not immediately attributable to B cell function, such as *cardiomyopathy* and *morning person*. This is

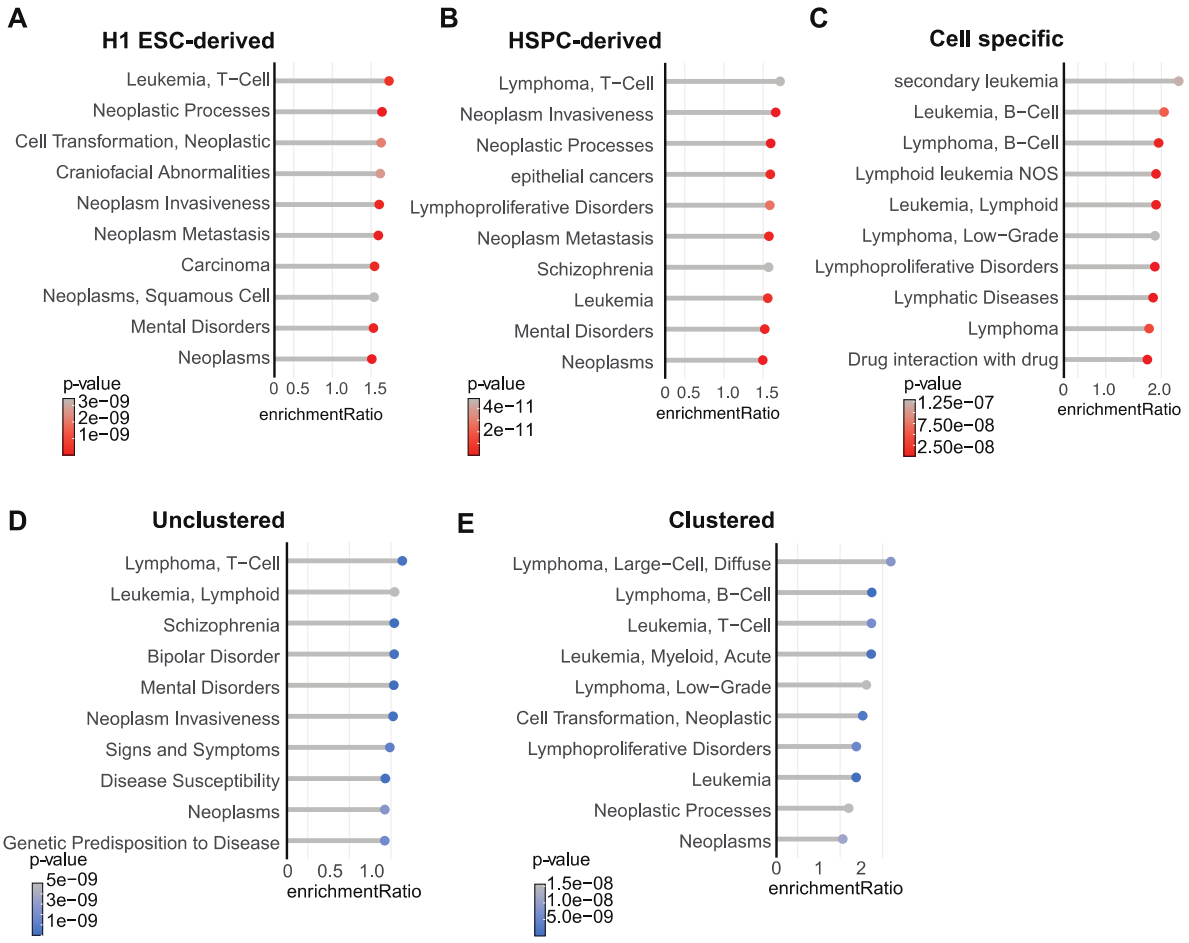
unsurprising due to the pleiotropy of gene regulation and the shared genetic architecture between many complex traits. However, HSPC-derived HMRs are enriched for genetic heritability of general hematopoietic traits including *white blood cell, platelet,* and *neutrophil counts*. In highly B cell-specific HMRs, we identify a notable enrichment of specific immune-related clinical traits and lab values, several of which achieve significance after multiple testing correction (p<Bonferroni, n=79). These observations hold true for S-LDSC analysis in H1 ESC-derived, and cell specific liver HMRs (**Fig 16**), reinforcing the notion that cell stage-derived HMRs are indicative of stage-relevant gene regulatory needs.

Figure 16. S-LDSC identifies Liver HMR annotation-specific trait enrichments.
Volcano-style plots of S-LDSC partitioned heritability results across 79 traits are shown for two liver HMR groups: H1 ESC-derived and cell-specific. HMRs are ordered by the developmentally distinct cell type in which they were established. Each HMR group was tested for enrichment of genetic heritability with a standard set of 98 base annotations against traits that include both clinical diseases as well as clinical lab values. Negative enrichment values were clipped to the lowest positive enrichment value for each row of plots (A: 0.02781896; B: 0.03787533). The size of each point represents the -$\log_{10}$*p*-value of the enrichment, and the color shows the $\log_{10}$enrichment value. Points with a *p*-value of ≤ 0.05 or an enrichment > 10 are labeled by their trait name where available.



HMRs stratified solely by clustering behavior also demonstrate heritability enrichment patterns associated with specific lymphoid traits (**Fig 15B**). In fact, compared to clustered HMRs, unclustered HMRs show no statistically significant trait enrichment above significance thresholds, suggesting that results in **Fig 15A** are powered predominantly by clustered HMRs. Accordingly, these trends were observed in gene-based disease enrichment analyses (disease ontology) applied to the same HMR groups analyzed by S-LDSC (**Fig 17**) (147, 148). For example, the top disease ontology enrichments for H1 ESC-derived HMRs include morphogenic ontologies such as *craniofacial abnormalities*, and the top ontologies for B cell-specific HMRs include multiple lymphoid- and leukemia-related ontologies, reflecting the biological state associated with each HMR group. Together with the partitioned heritability results, these data suggest clustered cell specific HMRs are both near lineage-specific genes and enrich for cell specific trait heritability over that of unclustered HMRs.

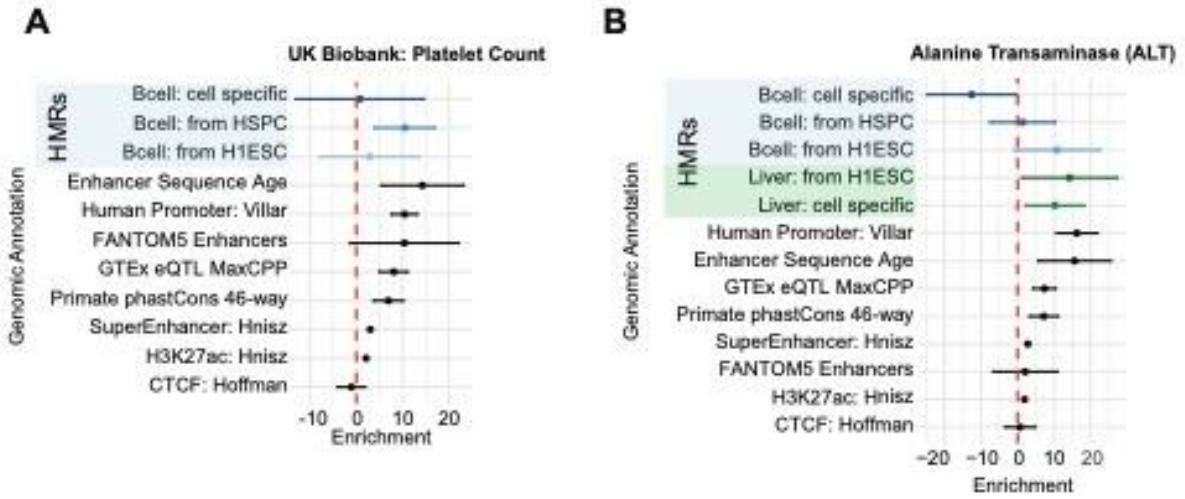
Figure 17. Disease ontology for developmentally specific and clustered B cell HMRs.
Lollipop plots show top ten disease ontology enrichments as analyzed through WebGestalt with default parameters. The *x*-axis shows enrichment ratios, and the *y*-axis displays disease ontologies sourced for the GLAD4U disease database (149). The *y*-axis is sorted by enrichment value. The color for each bar represents the *p*-value for that trait. Individual graphs show results from B cell HMR developmental and clustering groups: (A) H1 ESC-derived, (B) HSPC-derived, (C) cell-specific, (D) clustered, and (E) unclustered.



To better contextualize the partitioned heritability enrichment results from B cell data, we compared results against other known functional genomic feature annotations. We compared S-LDSC enrichment levels on a per-trait basis for B cell and liver HMR annotations (those from **Fig 15A and Fig 16)** and other functional genomic annotations (**Fig 15C-D, Fig 18A-B**). For both immune-related clinical lab values and disease traits, we observe increasingly stronger enrichment from H1 ESC-derived to HSPC-derived to B cell-specific HMRs. In contrast, both H1 ESC-derived and liver-specific HMRs show positive enrichment for ALT (alanine transaminase*)* compared to B cell HMRs, as expected. This shows that SNP-based trait enrichment is capable of distinguishing HMR patterns from both distant and highly related cell types. Across cell relevant traits, we observe SNP-based heritability enrichment values that surpass those of promoters, expression quantitative loci (eQTLs),

and histone marks of open chromatin (H3K27ac) often used to approximate active regulatory regions. Enrichment values associated with cell specific HMRs are comparable to those of FANTOM5 enhancers, supporting the notion that developmentally specific HMRs mark enhancers important for cell identity. Altogether, this analysis highlights the functional significance of different HMR patterns, all of which are enriched for heritability at or above the levels measured for other enhancer definitions. These results further indicate a quantitative relationship between HMR patterns and complex trait heritability. Thus, the stratification of HMRs by "sharedness" between cell types provides important contextual information to predict genome-to-trait relationships.

Figure 18. S-LDSC B cell by trait across genomic annotations.
Point and line pots of S-LDSC enrichent vlaues by annotation group per trait. The *x*-ais represents enrichment values, and the *y*-ais dispalys genomic annotations. Points show enirhcment point enstimates and lines display 95% confidence intervals. The red dotted line marks an enrichment score of 0. Annotation groups include popular enhancer-associated genomic annotatoins such as ancient human enhancer sequence age, FANTOM 5 enhancers, eQTLs, super-enhancers, and the H3K27ac histone mark. Genomic controls were also included, such as phastCons 46-way annotations as wella s promtoers and CTCF sites. The graphs include data from (A) develoop[metally dreived B cell HMRs. (B) This graph shows S-LDSC results for alanine transaminase. The data includes the annotations from (A) in addition to developmentally derived Liver HMRs.

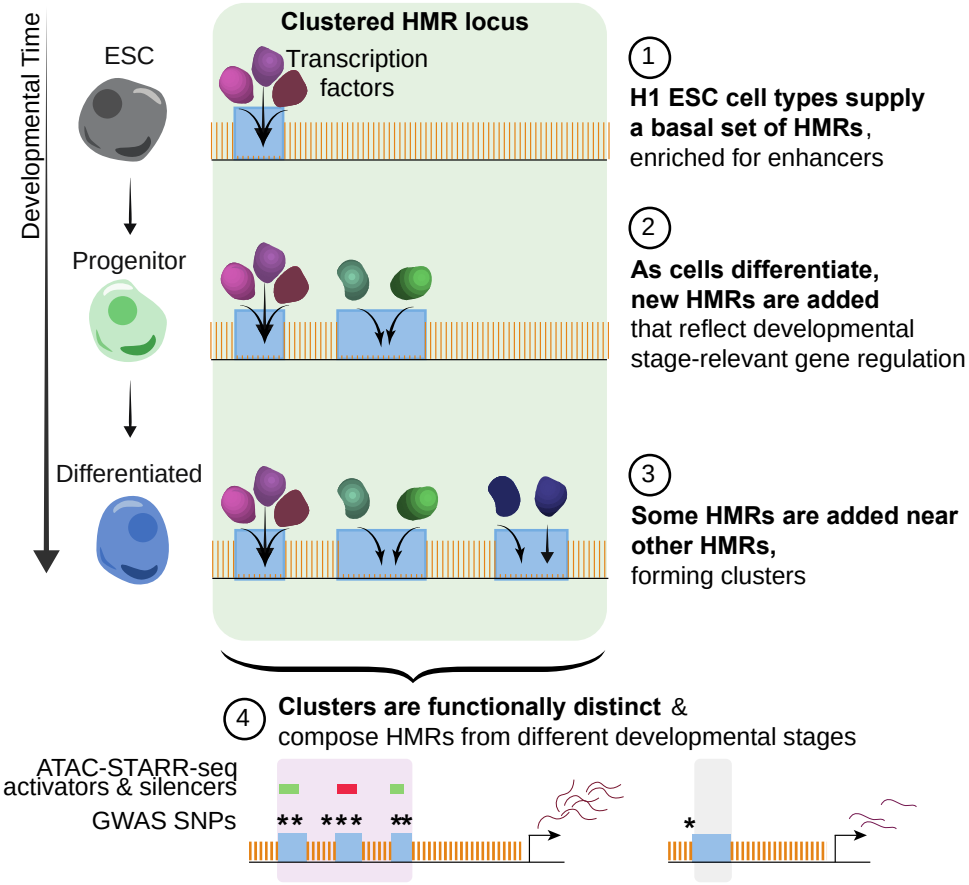


DISCUSSION

Here, we use comparative hypomethylation profiling to assess global hypomethylation patterns across cell types. This broader analysis reveals complex patterns of HMR establishment across a developmentally diverse dataset. By examining HMRs in a hematopoietic developmental context, we show that HMRs accumulate at distinct developmental stages and commonly persist through sequential lineage commitments.

These developmentally hypomethylated regions are associated with distinct, stage-appropriate transcription factors and gene pathways, leading to a model where H1 ESCs, with the fewest HMRs, present a basal set of HMRs to which additional regions are hypomethylated through development **(Fig 19).** In fact, about two-thirds of HMRs established in H1 ESCs remain in HMR datasets of differentiated cell types, highlighting their early establishment and continuous hypomethylation across time. Consequently, most (~3/4) HMRs in B cells were traced back to either H1 ESCs or HSPCs, indicating that the majority of HMRs are established at early cellular states. This further implies that biological differences between these cell types are driven by the minority population of differentially methylated HMRs. There are some exceptions to this general model, where a small subset of HMRs is "remethylated" between HSPCs and B cells. These regions are likely enhancers of genes involved in myeloid specification, as indicated by their retention in macrophage cells.

Figure 19. HMRs accumulate in clusters that record histories of cell development.
The conceptual model diagram summarizes the observations of HMR accumulation into clusters that feature different levels of methylation specificity.



Differentially methylated regions (DMRs) can be quantitatively identified and have been commonly used as a unit for studying DNA methylation (19, 150-153). However, our results demonstrate that consideration of HMRs that are shared among different degrees of developmentally related cell types can be highly informative for understanding the developmental history of the cell. The ability to distinguish unique HMRs between highly related cell types suggests that we can use combinatorial patterns of both shared and unique HMRs to distinguish or even predict cell types, although this remains to be tested.

We further show that new HMRs are preferentially established near existing HMRs, leading to the progressive enrichment of HMR clusters in differentiated hematopoietic cell types; however, it is unclear if these patterns extend to other developmental lineages. Notably, clustered HMRs compose about 1/3 of all HMRs in differentiated cells, compared to less than 1/6 in H1 ESCs, indicating clustering increases proportionally to developmental progression. These spatially correlated HMRs are enriched for unique stage-relevant gene ontologies, trait-associated genetic heritability, and ChromHMM annotations, implying distinct regulatory roles compared to their unclustered counterparts.

Previous investigations into enhancers describe subsets of clustering enhancers, including super-enhancers and hub enhancers (71, 104, 132). Super-enhancers that have been defined by H3K27 acetylation levels or by TF binding often consist of enhancer clusters. Clustering alone does not designate SEs and only a fraction of SEs is comprised of multiple enhancers units. We wanted to understand how many B cell HMR clusters also overlap super-enhancers that have been defined by ChIP-seq approaches. Our main conclusion from this analysis is that most super-enhancers also overlap clustered HMRs, but there are many more clustered HMRs than super-enhancers. One explanation for this broader phenomenon may be the finding that HMRs are often established near existing HMRs over developmental timescales. Thus, clustered HMRs can consist of regions representing both past and present enhancer activity (perhaps long after histone modifications and TF binding are lost). The establishment and maintenance of HMRs represents a unique characteristic of DNA hypomethylation compared to other more transient chromatin states. Another important consideration is that super-enhancers are defined by strength of TF binding or histone modification, which is measured on a continuous scale, whereas methylation is measured on an absolute scale.

We note that HMR clusters show patterns of hierarchical establishment that logically follow developmental paths. However, it is unclear if HMRs that persist through cell states remain

epigenetically active at later stages. Clusters may include a combination of active and decommissioned, inactive enhancers recorded in HMR patterns. Murine models of early development have highlighted contrasting dynamics between spatially and functionally related enhancers during exit from pluripotency, where some require re-methylation while others retain hypomethylation (62, 105, 115, 116, 154). These enhancers may serve to uphold cellular states during cell fate transitions. Our lab has also observed 'vestigial' enhancers on shorter timescales by applying ATAC-Me-seq to a differentiation time course, simultaneously measuring chromatin accessibility and DNA methylation (102); these analyses reveal a subset of regions that undergo chromatin closing while simultaneously maintaining hypomethylation levels. This is contrary to previous models of chromatin dynamics and DNA methylation that predicted methylation gain accompanies chromatin closing. These data suggest the uncoupling of chromatin accessibility and DNA methylation dynamics in a way that leads to the persistence of HMRs in chromatin inaccessible regions.

Partitioned heritability analysis of B cell HMRs established at three distinct developmental stages revealed enrichment of traits that reflect stage-relevant biology; in general, broadly shared HMRs were enriched for heritability of broader phenotypes while B cell-specific HMRs were enriched for lymphocyte-relevant traits. Each B cell HMR subset likely suffers from power limitations, representing between 3,187,775 and 9,228,469 bp, or as low as ~0.1% of the genome. Despite this limitation, we observe a remarkable correspondence between heritability enrichment and stage specific HMRs. This highlights the unique ability for hypo-methylation to capture information from multiple developmental timepoints; we find highly shared, lineage-shared, and cell specific heritability enrichment all within the methylome of a differentiated cell type. The genome-wide combination of stage-specific heritability signals within clusters implies the information is not only persistent through later cell stages, but also accumulated over time. The observation that DNA hypomethylation can persist through the closing of chromatin suggests that the use of H3K27ac to

identify putative enhancers precludes the observation of many HMRs, a subset of which forms hierarchical clusters that record cell developmental histories.

Our findings highlight DNA hypo-methylation as a unique epigenetic mark compared to common enhancer-associated histone marks. We highlight the unique accumulation of HMRs through developmental progression into clusters, enriched for stage-relevant SNP-based heritability. Through this process, epigenetic information can be maintained state to state. Thus, our results support that the methylome presents a historical documentation of developmental choices which could assist in the prioritization and interpretation of SNP data associated with clinical traits and diseases. These conclusions may further assist in understanding the complex role of the methylome in development and epigenetic gene regulation.

CONCLUSIONS

Here, we characterize HMR relationships both within and between developmentally diverse cell types to understand the functional significance of complex HMR patterns. We show that levels of HMR specificity across cell-types capture time point-specific branchpoints of development. Our analysis further reveals that HMRs form clusters in proximity to active genes that are important for cell identity. This is a wide-spread phenomenon and only a very small subset of HMR clusters is explained by overlapping super-enhancer annotations. Lastly, partitioned heritability revealed the functional significance of different HMR patterns linked to specific phenotypic outcomes and indicates a quantitative relationship between HMR patterns and complex trait heritability. Altogether, our findings reveal that HMRs can predict cellular phenotypes by providing genetically distinct historical records of a cell's journey through development, ultimately providing novel insights into how DNA *hypo*-methylation mediates genome function.

METHODS

HMR selection/exclusion dataset

DNA HMRs were obtained through the *MethBase* DNA Methylation trackhub from the UCSC Genome Browser, which references data processed through the *MethPipe* software for processing whole genome bisulfite sequencing data (108, 109). To achieve a high-confidence genome-wide methylation dataset, cell types were included based on a minimum coverage of 10x (155). This resulted in the selection of: *adrenal, fetal heart, fetal spinal cord, liver, macrophage, neutrophil,* and *T cell* from the NIH Roadmap Epigenomics Consortium (35); *H1 ESC* from Lister, et al. (92); and *B cell, neutrophil,* and *HSPC* (listed as "HSC" on the Genome Browser) from Hodges, et al. (32). As a primary cleaning step, to focus on non-coding HMRs, we removed promoter- and exon-overlapping HMRs (discussed below in section "Clustering annotation and percentage" and **Fig 10**). To do this, we combined RefSeq exon and protein-coding gene TSSs (-2000, +1000 bp) annotations to form an exclusion BED file. Next, we referenced this file to eliminate promoter- and exon-overlapping HMRs using the *intersect* function from the *Bedtools* package with option '-v' (107). Exclusion was defined by any basepair overlap. We required a minimum of 50bp for an HMR to be included in our analysis.

***CD27* multiple alignment with Hi-C**

Plots were generated in reference to the *hg19* genome build, showing the chromosome position: Chr 12: 6,522,500 – 6,575,000. We used a page width of 7, while HMR and methylation elements had a height of .3 and 1.0, respectively. The multiple alignment was constructed with the *plotgardener* R package (106), using the methylation and heatmap data represented in our HMR selection dataset. Hi-C interaction score data was visualized with the *plotHicTriangle*() function from the *plotgardener* R package (106). Contact matrices for both samples were plotted at 10 kb resolution.

HMR dendrogram

This analysis was performed using the per-HMR average CpG numerical matrix as composed in the methylation heatmap analysis. To perform hierarchical clustering, we used the *hclust*() function with the method "ward.D2". Colors were added using *Adobe Illustrator*.

Methylation heatmap

Heatmaps were generated in R with the package *pheatmap* (156). Numerical matrices representing per-HMR DNA methylation per cell type were used as input. These were generated in bash using methylation bigWig files from the *MethBase* DNA Methylation trackhub hosted on the UCSC Genome Browser (108, 109, 155). We used the *KentUtils* binary package to convert bigWig files to bedGraph files. bedGraph score columns were used to populate a numerical matrix representing the sample-population methylation proportion at individual HMRs in rows. The HMR consensus set used here represented all HMRs, created by concatenating HMR files from all cell types and using *Bedtools merge* to combine overlapping features. The HMRs from each cell type were filtered for a minimum length of 50 bp and against the list of RefSeq TSSs (-2000/+1000) and exons described above. Heatmaps were generated using R package: *pheatmap* using options: *kmeans_k* = 10, *cluster_cols* = FALSE, *cuttree_rows* = 10 (156). We also used the option "set.seed(86)" in R for reproducibility.

Transcription factor motif enrichment analysis

The *HOMER* perl package was used to calculate transcription factor motif enrichment (157). A background region was used to represent the merged HMR datasets of all cell types. Natural log transformed binomial p-values as reported by *HOMER* were used to rank motif enrichment output. *Scaled fold enrichment* was calculated by the quotient of two *HOMER* output values: [*%target/ %background]*. Top representative TFs are displayed in **Fig 3C**. All TFs shown represent the top TF by rank unless the top TFs were redundant. The second ranked TF is shown for the group, "Myeloid + HSPC," and the third ranked TF is shown for the group, "Differentiated." All data is represented in

Fig 11C to visualize TF enrichment differences between clustered and unclustered HMRs. Data visualization is scaled by TF to show relative cell specific enrichment. Graphing was performed in R using the *ggplot2* package (158).

k-means clustering gene ontology

Gene ontology was conducted using the web-based tool: GREAT (114). Specifically, GREAT takes BED files as input and assigns gene pairs using regulatory domains around gene TSSs (extending to the nearest gene's central domain up to a maximum extension distance). Here, we used the default gene annotation protocol from GREAT with a maximum extension of 1Mb. For input, we supplied BED files for each *k*-means cluster representing the HMRs in each group. Standard settings for maximum region-gene distance and gene assignment were used. Top results were downloaded from the web app using the "Shown ontology data as .tsv" selection. GREAT provided output for all k-means groups except for "Differentiated," as this group includes >20,000 HMRs and annotates to a large number of genes that prohibits the ability to detect gene ontology enrichment. Output files were filtered to exclude the top row before import to R. Top ranked binomial test *q*-value results are displayed as bar plots using *ggplot2* and *geom_bar()*.

Inter-HMR distances

To measure inter-HMR distances, we employed the *Bedtools closest* function with the '-io -d' options to calculate the distance from each HMR to the nearest HMR per cell type (107). Next, we extracted the output distance column to represent our observed distribution for graphing in R. To compare this distribution to random expectation, we used a script based on the process used for shuffling in Benton M.L., 2018 which uses *Bedtools closest* to calculate distances between shuffled non-coding HMRs per cell type (159, 160). The *Bedtools closest* function takes two input files. For this analysis, the input dataset is submitted twice. A region blacklist was used to exclude placement of HMRs

during the shuffle into coding space, defined by RefSeq TSSs and exons (161); this file was also used in the HMR annotation step. We iterated the random shuffle-closest procedure 10,000 times to create a null expectation of genomic positioning given random placement. Distances per shuffle-closest were summarized as means, yielding a distribution of average distances per shuffle. Distributions were plotted using the *ggplot2* R package. Inter-HMR distance values were filtered for those at or less than 500,000 bp to allow for better resolution of the density plot. Statistical significance between expected and observed inter-HMR distance values for each cell type were calculated using the *wilcox.test*() function in R. Statistical tests were computed on the list of inter-HMR distance values less than or equal to 500 kb.

Clustering annotation and percentage

To assess the prevalence of clustering per cell type, we utilized the same procedures outlined in **Fig 10**. Unclustered HMRs (**Fig 10A**) were defined as HMRs that are not within 6kb of any other HMR. We used *Bedtools merge* with the options '-c' and '-d 6000' to link BED regions and output constituent counts. We then use those that have a count of one and filter against RefSeq TSSs and exons. Because this excludes promoter and exon-proximal (within 6 kb) non-coding HMRs, we also perform a more inclusive unclustered HMR annotation by filtering HMRs by RefSeq TSSs and exons before performing a *Bedtools merge* step ('-c', '-d 6000') to identify isolated HMRs (**Fig 10B**), where *Bedtools merge* reports an input BED region that was not merged with any other HMR with a value of 1. By removing RefSeq TSS- and exon-overlapping HMRs before merging regions, we can find otherwise "unclustered" HMRs that are near a genic HMR. For **Fig 8C**, we subtract the total of our working unclustered HMR definition (**Fig 10A**) from this more inclusive definition to deduce the count of "TSS/exon-proximal" unclustered HMRs. To find clustered HMRs that do not cross the boundaries of TSSs and exons, we first use *Bedtools complement* to generate a BED file of regions that do not overlap the RefSeq regions. We then use *Bedtools intersect* with the '-c' and '-F 1.0' options to find a whitelist

set of regions that contain two or more (for identifying clusters with exactly 2 HMRs) or 3 or more HMRs. We use *Bedtools intersect* again with the '-lof' and '-F 1.0' options to produce a file where each row contains two BED coordinates: one for the whitelist region and one for the individual HMR. We then use a bash script to process this file with the purpose of linking HMR regions that are within 6 kb of each other that are within the same whitelist region (without passing a TSS or exon boundary). The output includes the linked end-to-end coordinates of clusters as well as the number of HMRs in each 6 kb-linked cluster. This can then be used to determine HMR clusters with exactly 2 (**Fig 10C**) or 3 or more HMRs (**Fig 10D**). To find individual clustered HMRs, we used *Bedtools intersect* with the original file as the '-a' file and merged cluster datasets as '-b' files. For the analyses outside of **Fig 8C**, the terms "clustered" and "clusters" refers to clusters of 3 or more HMRs. For **Fig 8D**, data was compiled and binned into clustered (3+) or unclustered HMRs. Denominators were defined as the total number of clustered and unclustered HMRs so that relative quantities in each cell type are visually comparable. Plots were generated with the *ggplot2* R package.

Sankey Diagram

HMR counts for each Sankey node and flow were determined using bash and the *Bedtools* suite. Nodes represent the total quantity of clustered and unclustered HMRs per cell type. Plots were generated in R using the package *networkD3* (162). To accurately represent the total quantity of HMRs per cell type, additional nodes were input and later processed with *Adobe Illustrator*.

Sankey gene ontology

Gene ontology was conducted using the web-based tool: GREAT (114), as with the *k*-means clustering gene ontology analysis. Here, we again used the default gene annotation protocol from GREAT. For a background file, we used the default "Whole genome" option. Standard settings for maximum region-gene distance and gene assignment were used. Top results were downloaded from the web app using

the “Shown ontology data as .tsv” selection. Output files were filtered to exclude the top row before import to R, and the preceding “#” is removed from the second row. Top results ranked by binomial *q*-value are displayed as a bar plot using *ggplot2* and *geom_bar*.

Super-enhancer annotation

GM12878 SEs were downloaded from the *Hnisz et al.* in hg19 as a BED file (of coordinates for both enhancers and super-enhancers) permitting comparison with clustered and all B cell HMR datasets using *Bedtools intersect (104)*. GM12878s are a well-studied Tier 1 ENCODE cell type derived from EBV-transformed B cells. SEs were selected from the “GM12878.bed” file. To use *eulerr*, input coordinates between groups must match; to accomplish this, we concatenated the GM12878 SE, B cell clustered, and B cell unclustered files before using *Bedtools* to sort and merge the BED file. We then used *Bedtools intersect* with the ‘-u’ option and the merged BED file as the ‘-a’ file to map each input BED file to the merged regions (representing a consensus list of HMRs). These were combined in R to generate a list of three BED files. Plotting was performed using the *R* package, *eulerr*, with the option ‘shape = “ellipse”’ to maintain proportionally sized ellipses.

ChromHMM annotation

A ChromHMM 15-state annotation file was downloaded from the UCSC Genome Browser in hg19 as assayed in the GM12878 cell line (155). Intersections were assessed using *Bedtools intersect* with the ‘-wo’ option and B cell HMRs as the ‘-a’ file with the ChromHMM BED file as the ‘-b’ file. Using R, HMR quantities per ChromHMM group were calculated as the number of HMRs that contain at least one instance of that ChromHMM group. Denominators for calculation proportions were 5974 and 17185 for B cell clustered and unclustered HMRs, respectively. Statistical testing was performed using the Z-test of proportions in R using *prop.test()*. Graphing was performed in *R* with the package, *ggplot2*.

ATAC-STARR-seq comparison

BED files for GM12878 ATAC-STARR-seq regulatory elements were obtained from Hansen & Hodges (134) (GSE181317). HMRs were converted to GRCh38 for comparison using *liftOver* (parameters: -bedPlus=3). We determined the number of overlaps between the datasets with Bedtools *intersect* (default parameters) piped to a line count command (wc -l). The proportion of overlapping HMRs was calculated as [#overlapping/#total] and then plotted with *ggplot2* in R. We performed a two-tailed, two-sample Z-test of proportions with the *prop.test()* function in R to obtain *p*-values.

Nearest-neighbor RNA-seq analyses

To determine if clustered HMRs are more likely to associate with active genes than unclustered HMRs, we measured the proportion of HMRs near "active" genes (TPM > 0) at different distances for the two HMR groups: HMR clusters that contain cell specific HMRs and unclustered cell specific HMRs. We defined coordinates for the clustered HMR region from end-to-end including all HMR constituents. To assign the nearest HMR-gene pair, we downloaded two RNA-seq datasets acquired from the ENCODEv3 release. Here, we elected to use data for GM12878s, a lymphoblastoid cell line, to match B cells as closely as possible; and we used the ENCODE Tier 1 HepG2 dataset to as a proxy for liver. We first isolated the ENSEMBL gene ID and TPM columns from each file before averaging between replicates for each cell type using the *tidyverse* package, merge(), and *rowMeans*() in R. We then used BioMart to convert ENSEMBL IDs to HUGO gene symbols to identify high-confidence protein-coding genes (163). To provide the highest conversion rate using BioMart, we truncated the version number from the ENSEMBL IDs. We performed the conversion using the *useMart*() function to establish search parameters with options: *biomart* = "ENSEMBL_MART_ENSEMBL," *host* = "grch37.ensembl.org," *path* = "/biomart/martservice," and *dataset* = "hsapiens_gene_ensembl." This was used in conjunction with the getBM() function requesting the output "attributes" of "hgnc_symbol," "strand," "chromosome_name," "start_position," and "end_position." We then filtered the output for rows that had a non-empty "hgnc_symbol" column value. This was then merged with

the dataframe described above with ENSEMBL ID and averaged TPM values. We used the strand information provided from BioMart to elect a TSS from either the "start_position" or "end_position," based on if the "strand" was "1" or "-1" respectively. This file was transformed into BED format using the TSS position to create a gene file with coordinate, gene ID, and TPM information.

To find the nearest active gene to HMR clusters and unclustered HMRs, we employed a strategy to first find a large pool of surrounding genes, before filtering out pairs that cross TAD boundaries and identifying the nearest gene. To do this, we assigned the surrounding gene landscape to each HMR by using *Bedtools closest* with the '-d' distance option as well as the '-k 100' option to output the nearest 100 genes to each HMR. We then filtered the list of HMR-TSS pairs for TAD crossing using reference TAD BED files, for "GM12878" and "Liver," as downloaded from the 3D Genome Browser (164). We used *Bedtools* intersect with the '-f 1.0' option to eliminate HMR-TSS pairs that are not fully within a TAD BED coordinate range. Using R, we filtered the resulting list to represent the nearest gene to each HMR. We then determined the quantity of HMR-gene pairs under each distance threshold (10, 25, 50, 75, 100, and 150 kb) for each HMR group, separately, by filtering the single nearest neighbor dataset by the HMR-TSS distance column and counting rows. We calculated the denominator of these proportions as the total amount of HMRs input to the analysis for each HMR group for each cell type. We used the *prop.test*() function in R to compare the HMR proportions between HMR clusters that contain cell specific HMRs and unclustered cell specific HMRs at each threshold value. Output was plotted using ggplot2, *geom_point*(), and *geom_line*().

To measure the transcriptional output differences associated with clustered or unclustered HMRs, we utilized the BED files of replicate-averaged TPM values and associated ENSEMBL IDs. We found the 2 nearest neighboring genes to each HMR using *Bedtools closest* with the '-d' distance option to output HMR-TSS distances and the '-k 2' option to limit output to the two nearest TSSs. We then filtered the list of HMR-TSS pairs for TAD crossing using reference TAD BED files, for "GM12878" and

"Liver," as downloaded from the 3D Genome Browser (164). We used *Bedtools* intersect with the '-f 1.0' option to eliminate HMR-TSS pairs not fully within a TAD BED coordinate range. We used R to eliminate gene redundancy within the clusters and unclustered datasets, separately. Statistical testing was performed using the *wilcox.test*() function in R. TPMs were plotted using *ggplot2* and *geom_boxplot*().

S-LDSC

Stratified LD-score regression was performed with *LDSC* using the appropriate python scripts distributed by the Price lab (https://github.com/bulik/ldsc) (88). Reference base annotation files were downloaded from the Price repository (Phase 3, version 2.2 annotations). We used the appropriate reference files coordinating with the 1000 Genomes baseline v2.2 scores and HapMap 3 SNPs (https://alkesgroup.broadinstitute.org/LDSCORE/). Summary statistics were collected from both the Price lab (https://alkesgroup.broadinstitute.org/LDSCORE/independent_ sumstats/) as well as the Neale lab heritability repository (https://nealelab.github.io/UKBB_ldsc/ index.html) (89). Traits were obtained based on their determined relevance to either broad cell-agnostic etiology or to biology specifically relating to B cells or Liver. This provided us the ability to determine specificity of results associated with varying cell specificity of HMRs. In total, we assessed 79 traits as described in **Table 5**. The primary *LDSC* program was run per HMR annotation per trait. Results for individual traits were tabularized per HMR annotation. Results were visualized using *ggplot* in R with the functions *geom_point* and *case_when* for conditional coloring. To determine B cell developmentally derived HMRs, we used *Bedtools intersect* to compare HMR files. H1 ESC-derived B cell HMRs were defined by B cell HMR coordinates and had to had overlap with HMRs from HSPC as well as H1 ESC, together. HSPC-derived B cell HMRs had to have overlap with HSPC HMRs but not H1 ESC HMRs. B cell-specific HMRs had to have no overlap with any HMRs from the collection of adrenal gland, liver, fetal heart, fetal spinal cord, H1 ESC, HSPC, macrophage, neutrophil, and T cell HMRs. In

the clustering analysis, all clustered or all unclustered HMRs were used. Liver HMRs were defined as H1 ESC-derived based on any overlap with H1 ESC HMRs. Cell specific Liver HMRs were also defined against the same comparative cell type collection used with B cell for this specific analysis. Annotations used to compare against HMR groups were selected from those included in the "baselineLF_v2.2.UKB.tar.gz" from the Price lab LD-score website. Annotations were selected for their relevance to enhancers; CTCF, a ubiquitous transcription factor, was included as a negative control for cell specific disease enrichment.

WebGestalt Gene Ontology Analysis

Developmentally grouped B cell HMR BED files, as used in the S-LDSC analysis, were used as input in addition to BED files for all B cell clustered or all B cell unclustered HMRs. GREAT Input was used to identify nearest neighbor genes in hg19 (114). We used the default gene assignment parameters under "Association rule settings" called "Basal plus extension," which in most cases replicates a two-nearest neighbor gene association strategy. In the web app, we downloaded the "Gene -> genomic region association table" file from the "genomic region-gene associations" page. Gene symbols were extracted from the GREAT input downloaded files using the first column. These were input into the WebGestalt web app to perform an over-representation analysis on the disease functional database, GLAD4U (147-149). For a reference gene set, we selected "genome protein-coding." Results were downloaded, and the enrichment values file was used to plot enrichment ratio values for top diseases in R using *ggplot2*.

Chapter III

DISCUSSION AND FUTURE DIRECTIONS

Discussion

This research aimed to examine the non-coding HMR patterns that arise from cross-cell type methylation profiling across both highly related and distant cell types. We collected whole-genome bisulfite sequencing data from multiple developmental stages, including H1 ESCs, fetal tissues, adrenal gland, liver, and representatives from the hematopoietic lineage (e.g., Hematopoietic stem and progenitor cells, macrophages, T cells, and B cells). Unsupervised clustering of over 126,000 HMRs across 9 cell types revealed distinct HMR groups that indicate a hierarchical establishment of HMRs through development; HMR groups reflect developmental stage specificity, supported by both transcription factor motif enrichment and gene ontology. By dissecting HMR patterns in a pseudo-time course from H1 ESCs to adult stem (HSPC) and differentiated cell types (Macrophage and B cell), we identify that most HMRs ($\sim$ 60%) are established in early stem cell stages and persist through subsequent developmental stages. These are adjoined by the accumulation of increasingly cell specific HMRs through differentiation; hierarchically established HMRs are formed nearby existing HMRs more often than expected, leading to $\sim$1/3 of HMRs existing in clusters ($\leq$ 6 kb between HMRs) in mature cell methylomes. We find clusters to be associated with increased enhancer activity. SNP-based partitioned heritability analysis reveals enrichment of complex trait genetic heritability in HMRs; furthermore, heritability for trait-relevant traits enriches positively with both HMR specificity and clustering.

Through unsupervised clustering, we identified groups of HMRs of varying degrees of specificity that reflect a developmental hierarchy. Hierarchical clustering groups reveal HMRs that are shared among all cell types; shared among cell types of a specific developmental stage (e.g. fetal tissue); shared among a specific lineage or sub-lineage; and cell specific. Previous studies have highlighted the cell specificity of methylomes and their ability to distinguish cell types from similar tissues. A study by Schultz, et al. annotated DMRs among the tissue and cell type groups: glands, mucosa, muscle, immune, fat, and epithelial (14). Hierarchical clustering applied to the set of 1,198,132 DMRs showed that differential methylation was sufficient to distinguish tissues and cell type groups. Hierarchical clustering of DMRs was as similarly discriminatory as RNA-seq data, emphasizing the specificity of the methylome.

HMRs are hierarchically acquired through development

Here, we characterize non-coding HMRs across the genome between diverse cell types (both highly related and distant), providing us granularity in assessing the degrees to which all non-coding HMRs are shared among cell types and tissues; in contrast, DMRs only capture those hypomethylated regions that differ between samples by a statistical threshold. DMRs are commonly identified through an assumption of a beta-binomial distribution to model methylation values (165); however, among a collection of samples from multiple tissue origins, the DMR definition lacks granularity in understanding the degree to which methylation profiles are shared among specific, related cell types. Thus, DMRs lack the resolution to interpret degrees of developmental specificity, whereas our whole-genome cross-cell type methylation profiling approach reveals HMR groups that we can identify as lineage specific (e.g. lymphoid specific, defined by the specific presence in

only B and T cells). These results underline the power of viewing the methylome through distinct HMR units and incorporating diverse cell types from various lineages and developmental timepoints to reveal specific branchpoints in cell fate specification.

We found novel insights regarding the developmental patterns of HMRs by analyzing methylomes within a pseudo-time course representing the hematopoietic lineage. We analyzed methylation data for the following collection of cell types: H1 ESC (pluripotent stem), HSPC (multipotent adult stem), macrophage (mature myeloid lineage), and B cell (mature lymphoid lineage); this combination of samples provided us with representatives of several sequential developmental stages. By measuring the total amount of HMRs across cell types, we found that HMRs tend to accumulate through differentiation in the hematopoietic lineage (other adult methylomes also show increased HMR counts compared to H1 ESC). Tracing the retention of existing HMRs through later stages revealed that a majority of HMRs are maintained in subsequent stages. The exception to this resides in the hematopoietic lymphoid lineage, where we observe a decrease in both total HMRs and HMRs retained from HSPC to B cell; this conforms with expectations from mouse work that indicates the need to remethylate enhancer regions during this transition to avoid a lineage imbalance favoring the myeloid lineage (115, 116).

The general retention of non-coding HMRs through developmental progression in whole-genome data conforms with observations in previous studies in mouse models. A study by Hon et al. utilized whole-genome methylation data from 17 cell types, describing a subset of DMRs that display an active chromatin state (as measured by H3K4me1 and H3K27ac) in mouse embryonic stem cells, but an inactive chromatin state in adult tissues (62). This

observation suggested that DNA methylation can record "vestigial" enhancers that were active at a previous developmental stage. This retention of DNA methylation contrasts with the emphasis on a hypomethylated genome in embryonic stem cells and subsequent methylation of promoter regions during exit from naïve pluripotency (45, 166).

Prior to the genome-wide scope afforded by whole-genome bisulfite sequencing, promoter methylation was measured by targeted bisulfite PCR amplification. While promoters of genes driving pluripotency have been shown to undergo methylation during cell fate transitions, this limited scope promoted the notion that pluripotency equated with vast hypomethylation and that methylation accompanied restriction of lineage determination. Our results contradict these assumptions as we show that H1 ESCs contain the fewest number of HMRs of any cell type in our analysis, and that most of the HMRs present in H1 ESCs are retained in all other cell types. Our analysis allowed us to build an alternative model whereby H1 ESCs feature the most restrictive hypomethylation profiles, and cell specification is accompanied by hypomethylation of stage-specific non-coding regulatory elements. A revised model that incorporates more recent observations from whole-genome methylation data could benefit the interpretation of DNA methylation studies as the field continues to gain whole-genome bisulfite datasets.

HMRs clusters offer a new model for clustered enhancers

Clustered enhancers have been studied through various types of functional genomic annotations and methodologies. Early studies identified specific locus control regions with multiple DNase I-hypersensitivity sites that affect gene transcription (69, 167-170). These highlighted the variable importance and cell-specific activity of constituent elements of

enhancer clusters, suggesting the influence of individual elements of an enhancer cluster is context specific. More recently, super-enhancers have been identified through ChIP-seq signal targeting the enhancer-associated marks: histone modifications (e.g. H3K27ac) and transcription factor binding (e.g. Pu1, Med1) (71, 104). As part of the annotation strategy, ChIP-seq peaks are linked (≤ 12.5 kb) together computationally to be analyzed as a singular unit, before looking for "exceptionally" high ChIP-seq signal for SE designation. It is important to note that because of this strategy, not all SEs contain constituent enhancers; ~ 15% are singular enhancers, and ~ 77% contain three or fewer constituents (72). However, the notion that all SEs represent clustered enhancers has been perpetuated by numerous studies that define SEs in language that suggests as much (104, 122, 171-175). As a popular annotation for clustered enhancers, we compared SEs defined in GM12878 (lymphoblastoid) cells by H3K27ac signal with our B cell HMR clusters. We find that while most SEs are captured by HMRs, there are many more HMR clusters than SEs, suggesting that clustered enhancers across the genome are underappreciated and understudied.

While SEs only represent those elements with exceptionally high ChIP-seq signal values, SEs do not capture all H3K27ac peaks. Nonetheless, we identify 27,505 unique HMR clusters in our comparison between H3K27ac-defined SEs in GM12878 cells (lymphoblastoid cells) and B cell HMR clusters, suggesting HMR clusters may be more pervasive than other clustered enhancer annotations (104)—and that subsets of HMRs do not overlap histone modifications indicative of open chromatin (e.g. H3K27ac). This corresponds with observations from our lab and others regarding the decoupling of chromatin accessibility and DNA methylation. Schultz, et al. identified 1,198,132 DMRs across 6 tissue groups, and further reported that 60.1% (719,837) were "novel" (14). These

were compared against putative enhancers defined by histone marks (176), suggesting that subsets of HMRs that are specific to cell types and lineages may exist outside of open chromatin histone marks.

Related to these ideas, our lab has developed ATAC-Me to simultaneously profile chromatin accessibility and methylation status through a protocol that combines use of Tn5 transposase to enrich for nuclease-free DNA as well as bisulfite conversion prior to sequencing (99, 102). When applied to a THP-1 monocyte-to-macrophage time course model (up to 72hr), dynamic chromatin accessibility changes were apparent; however, average methylation values within these opening and closing regions was largely constant. This was contrary to the idea that enhancer demethylation temporally accompanies chromatin opening. In contrast, Barnett et al. reveals a temporal decoupling of DNA methylation and chromatin accessibility, where changes in methylation are delayed compared to chromatin accessibility. Regions that underwent chromatin closing showed no statistical changes in methylation through this time course, suggesting that the underlying DNA sequence retained a hypomethylated state while transitioning to a closed chromatin state. Altogether, this argues that, while HMRs often coincide with chromatin accessibility, their existence is not dependent on open chromatin and may thus be more prevalent; this also supports a model whereby histone marks may reflect transient regulatory states, whereas DNA hypomethylation records both current and previous regulatory activity. This implies that DNA hypomethylation provides a unique view of genome function that links present and previous genome to phenome using information that may not be present in chromatin accessibility or active transcription factor binding data.

HMR subsets enrich for genetic heritability

S-LDSC applied to our HMR groups reveals that HMRs enrich for SNP-based partitioned genetic heritability for cell-relevant complex traits—both disease statuses and clinical lab values (88, 89). Additionally, by annotating B cell and Liver HMRs by the developmental stage (e.g. H1 ESC, HSPC, or mature cell type) at which they were established, we find that enrichment for cell-relevant heritability (e.g. *lymphocyte counts* measured in B cell HMRs) scales with the specificity of the HMRs as well as clustering status. This is consistent with our observations from unsupervised clustering combined with TF motif enrichment and GO, suggesting that HMRs established at a specific developmental stage overlap enhancers that are appropriately stage specific. Notably, in our S-LDSC analysis of B cell HMRs, we find B cell-specific HMRs to enrich heritability at levels above those for other enhancer annotations, including FANTOM5 enhancers (Fig. S11). FANTOM5 defines enhancers through balanced bidirectional capped transcripts and are considered to be a high-quality annotation for active enhancers (177). HMRs defined by clustering and their developmental stage of establishment—predictably enriched for stage-relevant trait SNP-based heritability—could provide a functional annotation resource to prioritize trait-associated SNPs, provided the appropriate cell types were assayed by whole-genome bisulfite sequencing. Linkage disequilibrium presents difficulty in differentiating causal SNPs from those SNPs that show association with a trait due to localization within the same LD block (178, 179). Overall, the strong enrichment for heritability with HMR subsets presents the ability to compare HMR information with other genetic layers (e.g. SNPs, chromatin accessibility, transcription factor motifs, or conserved sequence regions) to better understand the putative function of SNPs.

In conclusion, we have systematically dissected HMR patterns across diverse cell types representing a variety of lineages and developmental stages. Our results argue that all HMRs within the methylome of a cell type or tissue are informative in understanding and predicting cellular phenotypes. We show HMRs are highly predictive of cell identify and provide a unique epigenetic layer that captures genetically distinct records of cell's history through cell development.

Future Directions

Our data and that of others highlight the lineage and cell specificity of HMRs, storing information highly predictive of cell identity. A study by Koestler, et al. utilized methylation data from the Illumina Infinium HumanMethylation27 BeadArray, which measures 27,578 CpGs across 14,495 genes, to statistically train a model to predict the cell proportions from whole-blood samples (180). Complete blood cell counts were used as a ground truth to measure monocyte to lymphocyte ratios. While this analysis attempted to discriminate related mature blood cell types, and the BeadArray only measures a subset of all CpGs, the model was able to predict a proportion of .176 monocytes to .814 lymphocytes compared to .179 and .821 as measured by complete blood cell counts; these results display a root mean squared error of only 5-6%, underlining the power of DNA methylation to predict and differentiate related cell types. The comparison by Schultz, et al. between hierarchically clustered DMRs amongst 6 tissue groups to hierarchically clustered RNA-seq data showed a similar separation of tissue groups, further emphasizing the specificity and predictive power of HMRs (14).

The ability to not only distinguish mature cell types, but also track shared lineage histories among cell types, suggests a collection of HMRs across diverse cell types and developmental stages could be used to predict the identity of an unknown cell type. This may be particularly useful in identifying the cell origin of an unknown cancer clinically, leading to improved therapeutic approaches. Ideally, a predictive model using HMRs would incorporate a variety of cell types from many lineages, germ lines, and specialized tissues. Currently, publicly available whole-genome bisulfite sequencing data includes an array of hematopoietic cell types, given the availability of whole blood sampling. Other cell types are also available, though many are non-healthy, cancer cells, given the importance of methylation changes in defining cancer phenotypes. Besides a limited (but growing) breadth of cell type diversity in publicly available methylome collections, many cell types do not offer multiple WGBS datasets. This lack of depth of individual cell types and tissues precludes our ability to cross-reference methylomes within a cell type to generate a "high-confidence" HMR dataset; we are also unable to assign weights to individual HMRs to represent their estimated effect on transcription or other phenotypes, as one might perform with allelic counts in SNP-based transcriptome or polygenic risk score models. However, both clustering status and developmental specificity HMR information could be incorporated into a transcription predictive model to provide additional statistical weighting to individual SNPs. Nonetheless, the specificity afforded by HMRs presents a promising opportunity for cell identity prediction as future WGBS datasets are generated and shared.

Overall, our study provides a model to annotate and interpret the methylome in informative ways. The decoupling of DNA methylation from other enhancer-associated marks such as chromatin accessibility provides a unique epigenetic layer that records both active and

historical regulatory states. Future expansion of whole-genome methylome data will be useful in improving our interpretation of genetic variants, gene regulation, and methylation, itself.

REFERENCES

1. Reik W, Lewis A. Co-evolution of X-chromosome inactivation and imprinting in mammals. Nature Reviews Genetics. 2005;6(5):403-10.
2. Bourc'his Db, Xu G-L, Lin C-S, Bollman B, Bestor TH. Dnmt3L and the establishment of maternal genomic imprints. Science. 2001;294(5551):2536-9.
3. Hata K, Okano M, Lei H, Li E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. 2002.
4. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-57.
5. Jackson M, Krassowska A, Gilbert N, Chevassut T, Forrester L, Ansell J, et al. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Molecular and cellular biology. 2004;24(20):8862-71.
6. Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nature genetics. 2015;47(5):469-78.
7. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009;324(5929):930-5.
8. He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011;333(6047):1303-7.
9. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011;333(6047):1300-3.
10. Delhommeau F, Dupont S, Valle VD, James C, Trannoy S, Massé A, et al. Mutation in TET2 in myeloid cancers. New England Journal of Medicine. 2009;360(22):2289-301.
11. Abdel-Wahab O, Mullally A, Hedvat C, Garcia-Manero G, Patel J, Wadleigh M, et al. Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood, The Journal of the American Society of Hematology. 2009;114(1):144-7.
12. Antequera F, Bird A. CpG islands. DNA methylation. 1993:169-85.
13. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. nature. 2009;462(7271):315-22.
14. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212-6.
15. Holliday R, Grigg G. DNA methylation and mutation. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 1993;285(1):61-7.
16. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences. 2006;103(5):1412-7.
17. Bell JS, Vertino PM. Orphan CpG islands define a novel class of highly active enhancers. Epigenetics. 2017;12(6):449-64.

18. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. Journal of molecular biology. 1987;196(2):261-82.
19. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nature genetics. 2009;41(2):178-86.
20. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985;40(1):91-9.
21. Bird AP. CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics. 1987;3:342-7.
22. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Human genetics. 1988;78:151-5.
23. Tazi J, Bird A. Alternative chromatin structure at CpG islands. Cell. 1990;60(6):909-20.
24. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nature genetics. 2006;38(6):626-35.
25. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Current opinion in genetics & development. 1993;3(2):226-31.
26. Spruijt CG, Gnerlich F, Smits AH, Pfaffeneder T, Jansen PW, Bauer C, et al. Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives. Cell. 2013;152(5):1146-59.
27. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schübeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528(7583):575-9.
28. Hendrich B, Bird A. Identification and characterization of a family of mammalian methyl CpG-binding proteins. Genetics Research. 1998;72(1):59-72.
29. Baubec T, Ivánek R, Lienert F, Schübeler D. Methylation-dependent and-independent genomic targeting principles of the MBD protein family. Cell. 2013;153(2):480-92.
30. Gaston K, Fried M. CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. Nucleic acids research. 1995;23(6):901-9.
31. Choi JK. Contrasting chromatin organization of CpG islands and exons in the human genome. Genome biology. 2010;11(7):1-8.
32. Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, et al. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Molecular cell. 2011;44(1):17-28.
33. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, et al. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell. 2011;146(6):1029-41.
34. Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome research. 2013;23(10):1601-14.
35. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
36. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017;543(7643):72-7.

37. Holliday R, Pugh JE. DNA Modification Mechanisms and Gene Activity During Development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. Science. 1975;187(4173):226-32.
38. Compere SJ, Palmiter RD. DNA methylation controls the inducibility of the mouse metallothionein-I gene in lymphoid cells. Cell. 1981;25(1):233-40.
39. Busslinger M, Hurst J, Flavell R. DNA methylation and the regulation of globin gene expression. Cell. 1983;34(1):197-206.
40. Vardimon L, Kressmann A, Cedar H, Maechler M, Doerfler W. Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proceedings of the National Academy of Sciences. 1982;79(4):1073-7.
41. Stein R, Razin A, Cedar H. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proceedings of the National Academy of Sciences. 1982;79(11):3418-22.
42. Keshet I, Lieman-Hurwitz J, Cedar H. DNA methylation affects the formation of active chromatin. Cell. 1986;44(4):535-43.
43. Yisraeli J, Adelstein RS, Melloul D, Nudel U, Yaffe D, Cedar H. Muscle-specific activation of a methylated chimeric actin gene. Cell. 1986;46(3):409-16.
44. Kelley DE, Pollok BA, Atchison ML, Perry RP. The coupling between enhancer activity and hypomethylation of κ immunoglobulin genes is developmentally regulated. Molecular and cellular biology. 1988;8(2):930-7.
45. Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular cell. 2008;30(6):755-66.
46. Razin A, Cedar H. DNA methylation and gene expression. Microbiological reviews. 1991;55(3):451-8.
47. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics. 2006;38(12):1378-85.
48. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome research. 2013;23(3):555-67.
49. Li E, Beard C, Forster A, Bestor T, Jaenisch R, editors. DNA methylation, genomic imprinting, and mammalian development. Cold Spring Harbor symposia on quantitative biology; 1993: Citeseer.
50. Wilks AF, Cozens PJ, Mattaj IW, Jost J-P. Estrogen induces a demethylation at the 5'end region of the chicken vitellogenin gene. Proceedings of the National Academy of Sciences. 1982;79(14):4252-5.
51. Vedel M, Gomez-Garcia M, Sala M, Sala-Trepat JM. Changes in methylation pattern of albumin and α-fetoprotein genes in developing rat liver and neoplasia. Nucleic Acids Research. 1983;11(13):4335-54.
52. Gruenbaum Y, Stein R, Cedar H, Razin A. Methylation of CpG sequences in eukaryotic DNA. FEBS letters. 1981;124(1):67-71.
53. Gu J, Stevens M, Xing X, Li D, Zhang B, Payton JE, et al. Mapping of variable DNA methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. G3: Genes, Genomes, Genetics. 2016;6(4):973-86.

54. Rao X, Evans J, Chae H, Pilrose J, Kim S, Yan P, et al. CpG island shore methylation regulates caveolin-1 expression in breast cancer. Oncogene. 2013;32(38):4519-28.
55. Doi A, Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue-and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nature genetics. 2009;41(12):1350-3.
56. Kuo KC, McCune RA, Gehrke CW, Midgett R, Ehrlich M. Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. Nucleic acids research. 1980;8(20):4763-76.
57. Gama-Sosa MA, Midgett RM, Slagel VA, Githens S, Kuo KC, Gehrke CW, et al. Tissue-specific differences in DNA methylation in various mammals. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression. 1983;740(2):212-9.
58. Shapiro R, DeFate V, Welcher M. Deamination cytosine derivatives by bisulfite. Mechanism of the reaction. Journal of the American Chemical Society. 1974;96(3):906-12.
59. Susan JC, Harrison J, Paul CL, Frommer M. High sensitivity mapping of methylated cytosines. Nucleic acids research. 1994;22(15):2990-7.
60. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477-81.
61. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480(7378):490-5.
62. Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nature genetics. 2013;45(10):1198-206.
63. Magram J, Chada K, Costantini F. Developmental regulation of a cloned adult β-globin gene in transgenic mice. Nature. 1985;315(6017):338-40.
64. Kollias G, Wrighton N, Hurst J, Grosveld F. Regulated expression of human Aγ-, β-, and hybrid γβ-globin genes in transgenic mice: manipulation of the developmental expression patterns. Cell. 1986;46(1):89-94.
65. Kioussis D, Vanin E, DeLange T, Flavell R, Grosveld F. β-Globin gene inactivation by DNA translocation in γβ-thalassaemi. Nature. 1983;306(5944):662-6.
66. Driscoll MC, Dobkin CS, Alter BP. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5'beta-globin gene activation-region hypersensitive sites. Proceedings of the National Academy of Sciences. 1989;86(19):7470-4.
67. Forrester WC, Epner E, Driscoll MC, Enver T, Brice M, Papayannopoulou T, et al. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. Genes & development. 1990;4(10):1637-49.
68. Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell. 1987;51(6):975-85.
69. Li Q, Zhang M, Duan Z, Stamatoyannopoulos G. Structural analysis and mapping of DNase I hypersensitivity of HS5 of the β-globin locus control region. Genomics. 1999;61(2):183-93.
70. Pasquali L, Gaulton KJ, Rodríguez-Seguí SA, Mularoni L, Miguel-Escalada I, Akerman I, et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. Nature genetics. 2014;46(2):136-43.

71. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153(2):307-19.
72. Pott S, Lieb JD. What are super-enhancers? Nature genetics. 2015;47(1):8-12.
73. Parker SC, Stitzel ML, Taylor DL, Orozco JM, Erdos MR, Akiyama JA, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proceedings of the National Academy of Sciences. 2013;110(44):17921-6.
74. Quang DX, Erdos MR, Parker SC, Collins FS. Motif signatures in stretch enhancers are enriched for disease-associated genetic variants. Epigenetics & chromatin. 2015;8:1-14.
75. Shin HY, Willi M, Yoo KH, Zeng X, Wang C, Metser G, et al. Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nature genetics. 2016;48(8):904-11.
76. Scott TJ, Hansen TJ, McArthur E, Hodges E. Cross-tissue patterns of DNA hypomethylation reveal genetically distinct histories of cell development. BMC genomics. 2023;24(1):623.
77. Zhou HY, Katsman Y, Dhaliwal NK, Davidson S, Macpherson NN, Sakthidevi M, et al. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. Genes & development. 2014;28(24):2699-711.
78. Roden DM, Denny JC. Integrating electronic health record genotype and phenotype datasets to transform patient care. Clinical Pharmacology & Therapeutics. 2016;99(3):298-305.
79. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. Clinical and translational science. 2010;3(1):42-8.
80. Davis L. Psychiatric Genomics, Phenomics, and Ethics Research In A 270,000-Person Biobank (BioVU). European Neuropsychopharmacology. 2019;29:S739-S40.
81. Zhou W, Kanai M, Wu K-HH, Rasheed H, Tsuo K, Hirbo JB, et al. Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. Cell Genomics. 2022;2(10).
82. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009;106(23):9362-7.
83. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics. 2014;95(5):535-52.
84. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5.
85. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430(7001):743-7.
86. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature genetics. 2013;45(10):1238-43.
87. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9.

88. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature genetics. 2015;47(3):291-5.
89. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228-35.
90. Van der Ploeg L, Flavell R. DNA methylation in the human γδβ-globin locus in erythroid and nonerythroid tissues. Cell. 1980;19(4):947-58.
91. Kunnath L, Locker J. Characterization of DNA methylation in the rat. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression. 1982;699(3):264-71.
92. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315-22.
93. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics. 2007;39(4):457-66.
94. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, et al. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. PLoS genetics. 2007;3(10):e181.
95. Illingworth R, Kerr A, DeSousa D, Jørgensen H, Ellis P, Stalker J, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS biology. 2008;6(1):e22.
96. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766-70.
97. Portela A, Esteller M. Epigenetic modifications and human disease. Nature biotechnology. 2010;28(10):1057-68.
98. Bock C, Beerman I, Lien W-H, Smith ZD, Gu H, Boyle P, et al. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Molecular cell. 2012;47(4):633-47.
99. Barnett KR, Decato BE, Scott TJ, Hansen TJ, Chen B, Attalla J, et al. ATAC-Me captures prolonged DNA methylation of dynamic chromatin accessibility loci during cell fate transitions. Molecular cell. 2020;77(6):1350-64. e6.
100. He Y, Hariharan M, Gorkin DU, Dickel DE, Luo C, Castanon RG, et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature. 2020;583(7818):752-9.
101. Razin A, Szyf M. DNA methylation patterns formation and function. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression. 1984;782(4):331-42.
102. Guerin LN, Barnett KR, Hodges E. Dual detection of chromatin accessibility and DNA methylation using ATAC-Me. Nature Protocols. 2021;16(12):5377-97.
103. Dos Santos CO, Dolzhenko E, Hodges E, Smith AD, Hannon GJ. An epigenetic memory of pregnancy in the mouse mammary gland. Cell reports. 2015;11(7):1102-9.
104. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013;155(4):934-47.
105. Bell E, Curry EW, Megchelenbrink W, Jouneau L, Brochard V, Tomaz RA, et al. Dynamic CpG methylation delineates subregions within super-enhancers selectively decommissioned at the exit from naive pluripotency. Nature communications. 2020;11(1):1-16.

106. Kramer NE, Davis ES, Wenger CD, Deoudes EM, Parker SM, Love MI, et al. Plotgardener: Cultivating precise multi-panel figures in R. Bioinformatics. 2022;38(7):2042-5.
107. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
108. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PloS one. 2013;8(12):e81148.
109. Song Q, Decato B, Kessler M, Fang F, Qu J, Garvin T, et al. The Smithlab DNA Methylation Data Analysis Pipeline (MethPipe). 2021.
110. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194-211.
111. Rodda DJ, Chew J-L, Lim L-H, Loh Y-H, Wang B, Ng H-H, et al. Transcriptional regulation of nanog by OCT4 and SOX2. Journal of Biological Chemistry. 2005;280(26):24731-7.
112. Wedel A, Lömsziegler-Heitbrock H. The C/EBP family of transcription factors. Immunobiology. 1995;193(2-4):171-85.
113. Åkerblad P, Sigvardsson M. Early B cell factor is an activator of the B lymphoid kinase promoter in early B cell development. The Journal of Immunology. 1999;163(10):5453-61.
114. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology. 2010;28(5):495-501.
115. Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010;467(7313):338-42.
116. Bröske A-M, Vockentanz L, Kharazi S, Huska MR, Mancini E, Scheller M, et al. DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nature genetics. 2009;41(11):1207-15.
117. Izzo F, Lee SC, Poran A, Chaligne R, Gaiti F, Gross B, et al. DNA methylation disruption reshapes the hematopoietic differentiation landscape. Nature genetics. 2020;52(4):378-87.
118. Hendriks J, Gravestein LA, Tesselaar K, van Lier RA, Schumacher TN, Borst J. CD27 is required for generation and long-term maintenance of T cell immunity. Nature immunology. 2000;1(5):433-40.
119. Agematsu K, Hokibara S, Nagumo H, Komiyama A. CD27: a memory B-cell marker. Immunology today. 2000;21(5):204-6.
120. Lens SM, Tesselaar K, van Oers MH, van Lier RA, editors. Control of lymphocyte function through CD27–CD70 interactions. Seminars in immunology; 1998: Elsevier.
121. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell. 2013;153(2):320-34.
122. Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. Science. 2018;361(6400):eaar3958.
123. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SC, Erdos MR, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. Nature. 2015;520(7548):558-62.
124. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215-6.

125. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nature protocols. 2017;12(12):2478-92.
126. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics. 2007;39(3):311-8.
127. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010;107(50):21931-6.
128. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009;457(7231):854-8.
129. Capra JA. Extrapolating histone marks across developmental stages, tissues, and species: an enhancer prediction case study. BMC genomics. 2015;16(1):1-9.
130. Hong J-W, Hendrix DA, Levine MS. Shadow enhancers as a source of evolutionary novelty. Science. 2008;321(5894):1314-.
131. Cannavò E, Khoueiry P, Garfield DA, Geeleher P, Zichner T, Gustafson EH, et al. Shadow enhancers are pervasive features of developmental regulatory networks. Current Biology. 2016;26(1):38-51.
132. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nature communications. 2018;9(1):1-12.
133. Fraser P, Pruzina S, Antoniou M, Grosveld F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. Genes & development. 1993;7(1):106-13.
134. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor–bound activators and silencers within chromatin-accessible regions of the human genome. Genome Research. 2022;32(8):1529-41.
135. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nature communications. 2018;9(1):1-15.
136. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology. 2011;9(4):e1001046.
137. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.
138. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nature genetics. 2019;51(9):1380-8.
139. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.
140. Roederer M, Quaye L, Mangino M, Beddall MH, Mahnke Y, Chattopadhyay P, et al. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. Cell. 2015;161(2):387-403.
141. Chun S, Casparino A, Patsopoulos NA, Croteau-Chonka DC, Raby BA, De Jager PL, et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat Genet. 2017.

142. Guo MH, Nandakumar SK, Ulirsch JC, Zekavat SM, Buenrostro JD, Natarajan P, et al. Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc Natl Acad Sci U S A. 2017;114(3):E327-E36.
143. Vockley CM, Barrera A, Reddy TE. Decoding the role of regulatory element polymorphisms in complex disease. Curr Opin Genet Dev. 2017;43:38-45.
144. Koues OI, Kowalewski RA, Chang LW, Pyfrom SC, Schmidt JA, Luo H, et al. Enhancer sequence variants and transcription-factor deregulation synergize to construct pathogenic regulatory circuits in B-cell lymphoma. Immunity. 2015;42(1):186-98.
145. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell. 2016;167(5):1398-414.e24.
146. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518(7539):337-43.
147. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic acids research. 2005;33(suppl_2):W741-W8.
148. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research. 2019;47(W1):W199-W205.
149. Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. BMC genomics. 2012;13(8):1-12.
150. Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, Herb B, et al. Differential methylation of tissue-and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nature genetics. 2009;41(12):1350-3.
151. Frigola J, Song J, Stirzaker C, Hinshelwood RA, Peinado MA, Clark SJ. Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. Nature genetics. 2006;38(5):540-9.
152. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, Glass D, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS genetics. 2012;8(4):e1002629.
153. Hotta K, Kitamoto A, Kitamoto T, Ogawa Y, Honda Y, Kessoku T, et al. Identification of differentially methylated region (DMR) networks associated with progression of nonalcoholic fatty liver disease. Scientific reports. 2018;8(1):1-11.
154. Schulz M, Teissandier A, de la Mata E, Armand M, Iranzo J, El Marjou F, et al. DNA methylation restricts coordinated germline and neural fates in embryonic stem cell differentiation. bioRxiv. 2022:2022.10.22.513040.
155. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, et al. The UCSC genome browser database. Nucleic acids research. 2003;31(1):51-4.
156. Kolde R. Pheatmap: pretty heatmaps. R package version. 2012;1(2):726.
157. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell. 2010;38(4):576-89.
158. Wickham H. Package 'ggplot2': elegant graphics for data analysis. Springer-Verlag New York doi. 2016;10:978-0.
159. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer maps differ significantly in genomic distribution, evolution, and function. BioRxiv. 2017:176610.

160. Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. Bmc Genomics. 2019;20(1):1-22.
161. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016;44(D1):D733-D45.
162. Allaire J, Ellis P, Gandrud C, Kuo K, Lewis B, Owen J, et al. Package 'networkD3'. D3 JavaScript network graphs from R. 2017.
163. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439-40.
164. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome biology. 2018;19(1):1-12.
165. Chen D-P, Lin Y-C, Fann CS. Methods for identifying differentially methylated regions for sequence-and array-based data. Briefings in functional genomics. 2016;15(6):485-90.
166. Lee HJ, Hore TA, Reik W. Reprogramming the methylome: erasing memory and creating diversity. Cell stem cell. 2014;14(6):710-9.
167. Bender M, Bulger M, Close J, Groudine M. β-globin gene switching and DNase I sensitivity of the endogenous β-globin locus in mice do not require the locus control region. Molecular cell. 2000;5(2):387-93.
168. Chen H, Lowrey CH, Stamatoyannopoulos G. Analysis of enhancer function of the HS-40 core sequence of the human α-globin cluster. Nucleic acids research. 1997;25(14):2917-22.
169. Carter D, Chakalova L, Osborne CS, Dai Y-f, Fraser P. Long-range chromatin regulatory interactions in vivo. Nature genetics. 2002;32(4):623-6.
170. Tolhuis B, Palstra R-J, Splinter E, Grosveld F, De Laat W. Looping and interaction between hypersensitive sites in the active β-globin locus. Molecular cell. 2002;10(6):1453-65.
171. Khan A, Mathelier A, Zhang X. Super-enhancers are transcriptionally more active and cell type-specific than stretch enhancers. Epigenetics. 2018;13(9):910-22.
172. Blayney J, Francis H, Camellato B, Mitchell L, Stolper R, Boeke J, et al. Super-enhancers require a combination of classical enhancers and novel facilitator elements to drive high levels of gene expression. bioRxiv. 2022:2022.06. 20.496856.
173. Thandapani P. Super-enhancers in cancer. Pharmacology & therapeutics. 2019;199:129-38.
174. Yan B, Wang C, Chakravorty S, Zhang Z, Kadadi SD, Zhuang Y, et al. A comprehensive single cell data analysis of lymphoblastoid cells reveals the role of super-enhancers in maintaining EBV latency. Journal of Medical Virology. 2023;95(1):e28362.
175. Kai Y, Li BE, Zhu M, Li GY, Chen F, Han Y, et al. Mapping the evolving landscape of super-enhancers during cell differentiation. Genome Biology. 2021;22:1-21.
176. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 2015;518(7539):350-4.
177. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.

178. Aissani B. Confounding by linkage disequilibrium. Journal of human genetics. 2014;59(2):110-5.
179. Weir B. Linkage disequilibrium and association mapping. Annu Rev Genomics Hum Genet. 2008;9:129-42.
180. Koestler DC, Christensen BC, Karagas MR, Marsit CJ, Langevin SM, Kelsey KT, et al. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. Epigenetics. 2013;8(8):816-26.