Molecular Cell, *Volume 44*

Supplemental Information

Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment

Emily Hodges, Antoine Molaro, Camila O. Dos Santos, Pramod Thekkat, Qiang Song, Philip Uren, Jin Park, Jason Butler, Shahin Rafii, W. Richard McCombie, Andrew D. Smith, and Gregory J. Hannon

Contents

1 Supplemental Experimental Procedures

This supplement contains descriptions of many computational analysis methods used in the paper. A substantial portion of the methods have been used in Molaro et al. (2011) and therefore will not be described again here. Those methods include the basic pipeline for constructing the methylomes, the methodology for identifying HMRs, and the algorithm for measuring enrichment of one set of genomic intervals relative to another. We refer the reader to Molaro et al. (2011) for these details.

1.1 Hypomethylated regions

The method for identifying hypomethylated regions (HMRs) is described in Molaro et al. (2011) (in press). Briefly, HMRs were identified using a hidden Markov model (HMM) with 2 states: one for high methylation and one for low methylation. The data given to the HMM was the counts of methylated and unmethylated reads mapping over each CpG. These pairs of counts were modeled using a Beta-Binomial distribution. A cutoff for the minimum sum of posterior scores through an identified HMR was obtained by randomly permuting CpG sites and obtaining a size distribution for HMRs, and taking only those HMRs reaching the upper 1% of the scores obtained in the distribution obtained by randomization.

1.2 Differential methylation

Our general strategy for identifying differential methylation between two methylomes is to first calculate a differential methylation score for each individual CpG. Once the score has been calculated, we identify differentially methylated regions (DMRs) based on this score.

Our single-CpG differential methylation score is the probability that the CpG is methylated at a higher frequency in one methylome than the other. For CpG i, let m_i^a and u_i^a denote the number of methylated and unmethylated reads, respectively, in condition a. We assume p_i^a is the probability of methylation at CpG i in methylome a and that $p_i^a \sim$Beta(m_i^a, u_i^a). Given the observations of methylation at CpG i in conditions a and b, then we can use the exact formula for

$$\Pr(p_i^a > p_i^b) = f(m_i^a, u_i^a, m_i^b, u_i^b),$$

where the function f is as described by Altham (1969):

$$f(m_i^a, u_i^a, m_i^b, u_i^b) = \sum_{k=\max(m_i^b - u_i^a, 0)}^{m_i^b - 1} \frac{\binom{m_i^b + u_i^b - 1}{k} \binom{m_i^a + u_i^a - 1}{m_i^a + u_i^a - 1 - k}}{\binom{m_i^a + u_i^a + m_i^b + u_i^b - 2}{m_i^a + m_i^b - 1}}$$

This probability is symmetric and continuous, so for two methylomes a and b, we calculate either $\Pr(p_i^a > p_i^b) = 1 - \Pr(p_i^b > p_i^a)$.

Our algorithm for identifying differentially methylated regions consists of the following criteria:

- First the CpGs are partitioned into blocks such that no two consecutive CpGs is more than 500bp apart.
- Within each block, the differential methylation probabilities are smoothed using an Epanechnikov kernel with a bandwidth of 10 bases (not CpGs).
- After smoothing, we use a cutoff of 0.75 to identify peaks. For the opposite direction (since the scores are symmetric for the two methylomes being compared) we use 0.25.
- The set of candidate DMRs based on the cutoff are screened in two ways: (1) they must contain at least 10 CpGs, and (2) they must be at least 200bp in size.

We used randomization to indicate the cutoffs mentioned above. Briefly, the CpGs were randomly permuted within each block 100 times. For a given DMR size cutoff (both in terms of bases and CpGs) we obtain the number of DMRs identified, and compare this with the number identified in the real data. This provides a false discovery rate. We applied this method in several comparisons using an FDR of 0.05 to arrive at the cutoffs.

In Figure 3C we used differential methylation to correlate with differential expression. In this case we obtained the differential methylation probability $p_i(a, b) = \Pr(p_i^a > p_i^b)$ and then calculated the log-odds as $\log(p_i/(1 - p_i))$.

1.3 Measuring transcription factor binding site enrichment

In Figure 4B we measured transcription factor binding site (TFBS) enrichment inside the intergenic DMRs between neutrophils and B cells. The intergenic DMRs were selected for analysis using the following criteria:

1. Intergenic: residing at least 10Kbp from the nearest refGene.
2. Non-repeat: the DMRs must not overlap any annotated repeats, as downloaded through the UCSC Table Browser.
3. Non-CGI: the DMRs must not overlap an annotated CGI. The reason is that strong dinucleotide bias skews the motifs.

After applying these criteria, we ended with 1505 DMRs for N<B and 1175 for B<N. Before analyzing the sequences of these regions, they were expanded or contracted relative to their centers so that each sequence analyzed was 1Kbp in length.

The motif set used was a combination of known motifs from the JASPAR(Vlieghe et al., 2006) and TRANSFAC(Matys et al., 2006) databases. The total number of motifs in this data set was 775. For each motif we designated a family based on some shared binding property (*e.g.* motifs similar to CAGCTG were assigned the "EBOX2" class; motifs for all ETS family member TFs were designated "ETS").

We measured enrichment in the N<B DMRs relative to the B<N DMRs using the MOTIFCLASS program from the CREAD package(Smith et al., 2006). Briefly, MOTIFCLASS identifies the top scoring match

to each motif in each sequence from a foreground and background sequence set. We used the "binomial p-value" setting MOTIFCLASS, which for a given match score cutoff, calculates a p-value for enrichment in the foreground relative to the background from a binomial distribution. Let n be the total number of sequences with a match for the motif above a given cutoff, let k be the number of those from the foreground sequence set, and let p be the number of sequences in the foreground sequence set divided by the sum of the number of sequences in the foreground and the background. Then the p-value is for $\mathrm{Bin}(k; p, n)$. Finally, MOTIFCLASS optimizes the match score cutoff relative to this p-value. This procedure was applied once with N<B as foreground, and B<N as background; then the foreground and backgrounds were swapped and MOTIFCLASS was applied again. The p-values were not corrected for multiple hypothesis testing, as they were simply used to rank motifs; what is important in our results is the identity of the motifs landing in the top 20 for each DMR set, from among the 775 motifs evaluated in each case.

The method used to match motifs in sequences and identify the greatest match scores was described in (Hertz & Stormo, 1999) and implemented in the STORM program, also part of the CREAD package(Smith et al., 2006).

1.4 RNA-seq data processing

We used the RefSeq transcriptome as downloaded through the UCSC Table Browser (Karolchik et al., 2004). We mapped reads in two stages, first to sequences constructed from all RefSeq exons (with overlapping exons collapsed), and then to all possible junctions formed from all pairs of exons for the same gene. Mapping was done with RMAP (Smith et al., 2009) allowing up to 3 mismatches in 36 bases. Reads mapping ambiguously (including mapping to an exon and a junction) were discarded. For each RefSeq transcript we counted the number of reads whose mapping location was inside the transcript's exons (so a read can be counted for two transcripts, as long as the location is unique) or through one of the transcript's junctions. RPKM calculations discarded duplicate reads and corrected gene size for deadzones (portions of the transcripts to which no read can map uniquely).

Differential expression between two cells was computed using a 2×2 contingency table and chi-squared statistics or Fisher's exact test to obtain a p-value for differential expression. Briefly, the contingency tables contained, for each gene, the counts of reads inside the gene and outside the gene, for both cell types. We used Bonferroni correction for the p-values, and the remaining genes were called differentially expressed. These genes were then ranked based on RPKM ratios.

1.5 Cross species conservation

Cross species conservation was measured using phastCons scores as downloaded from the UCSC Genome Browser FTP server. These scores are posterior probabilities from a phylogenetic HMM (Siepel et al., 2005).

Supplemental References

Altham P (1969) Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *Journal of the Royal Statistical Society. Series B (Methodological)* 31:261–269.

Hertz G, Stormo G (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15:563–577.

Karolchik D, Hinrichs A, Furey T, Roskin K, Sugnet C, Haussler D, Kent W (2004) The UCSC Table Browser data retrieval tool. *Nucleic acids research* 32:D493.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34:D108–10.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15:1034–1050.

Smith AD, Sumazin P, Xuan Z, Zhang MQ (2006) DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc. Natl. Acad. Sci. USA* 103:6275–6280.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B (2006) A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 34:D95–7.

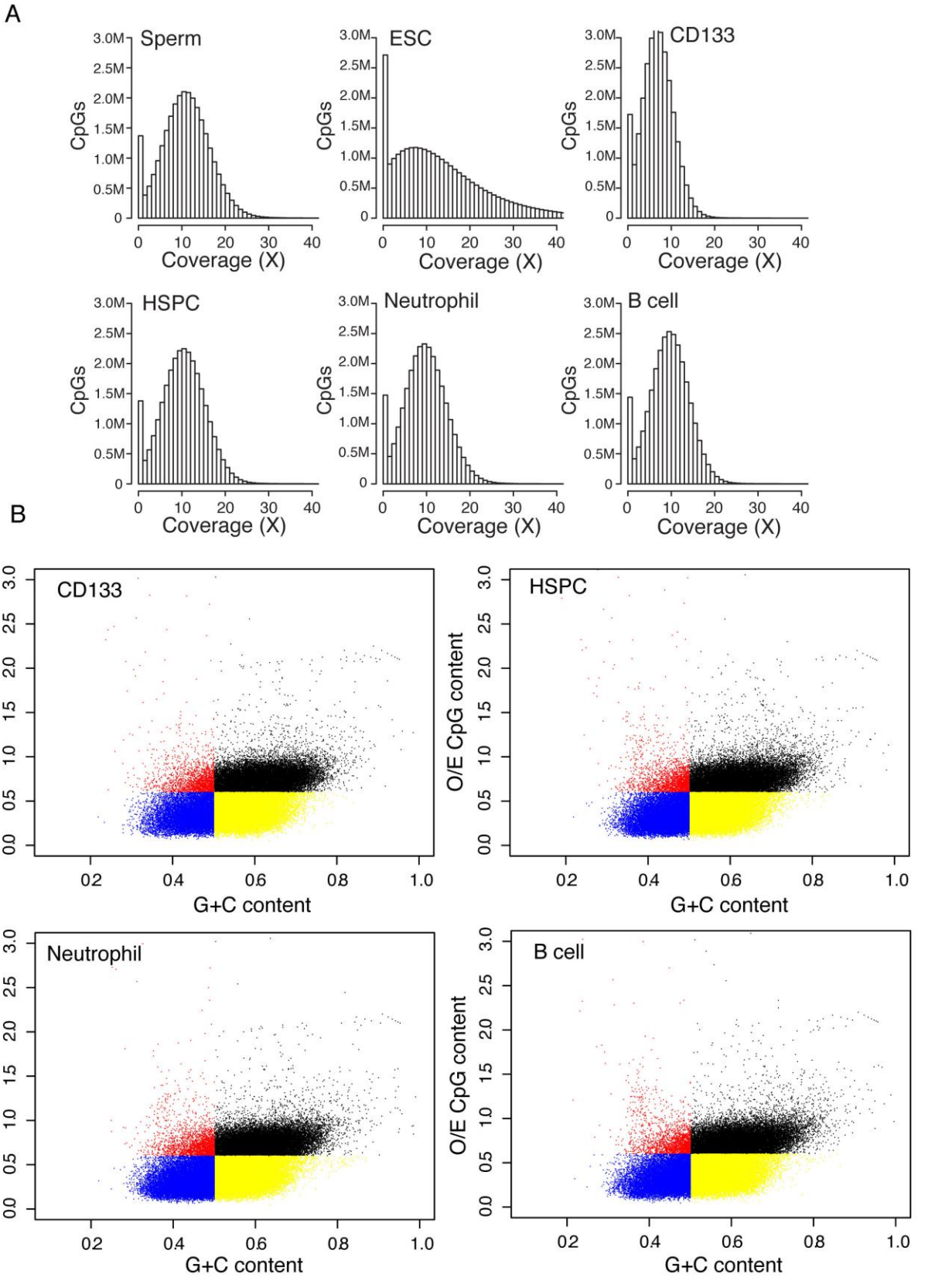

A
Sperm
ESC
CD133
HSPC
Neutrophil
B cell
CpGs
Coverage (X)
B
CD133
HSPC
Neutrophil
B cell
O/E CpG content
G+C content

Figure S1. CpG Mapping Coverage and HMR Characteristics, Related to Figure 1
(A) Distribution of read coverage for all CpG sites in the genome. Data is shown for human sperm, ESCs, CD133+ cord blood (stem cells), HSCs from peripheral blood, B cells and neutrophils. **(B)** G/C content and observed/expected CpG content for HMRs for each of the 6 cell types studied. Each HMR is a point, and colors indicate whether the HMR satisfies one or both of the sequence-based criteria described by Gardiner-Garden and Frommer and employed by UCSC to annotate CGIs genome-wide.

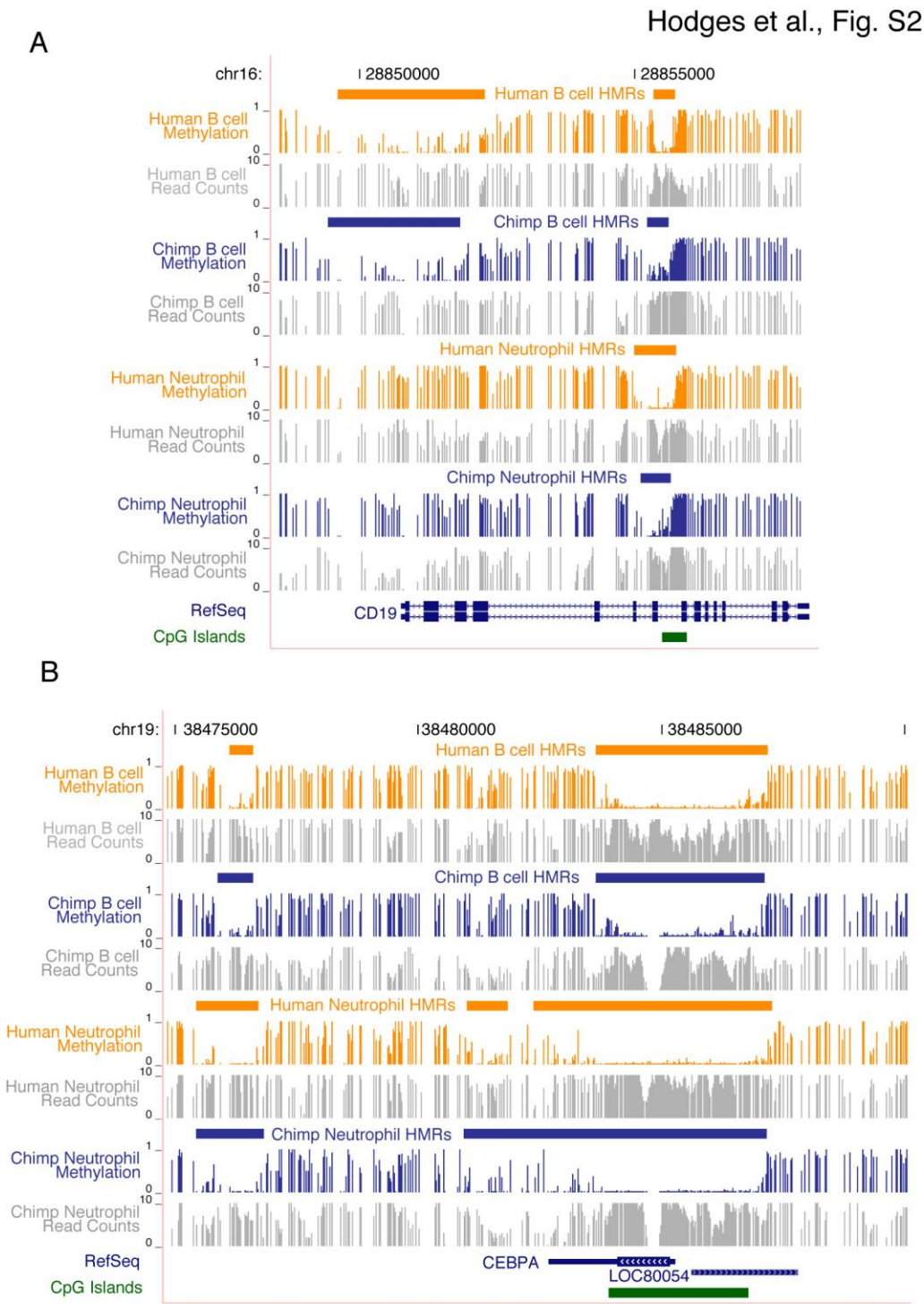


Figure S2. HMR Profiles Are Conserved between Chimp and Human, Related to Figure 1
Genome browser tracks depict methylation profiles across a lymphoid **(A)** and myeloid **(B)** specific locus in chimp and human blood cells. Methylation frequencies, ranging between 0 and 1, of unique reads covering individual CpG sites are shown in gray with identified hypomethylated regions (HMRs) indicated by orange bars. UCSC predicted/annotated CpG islands (green bars) as well as HMM-based CpG islands (blue bars) are also displayed. Numbers (top) indicate base position along the chromosome.

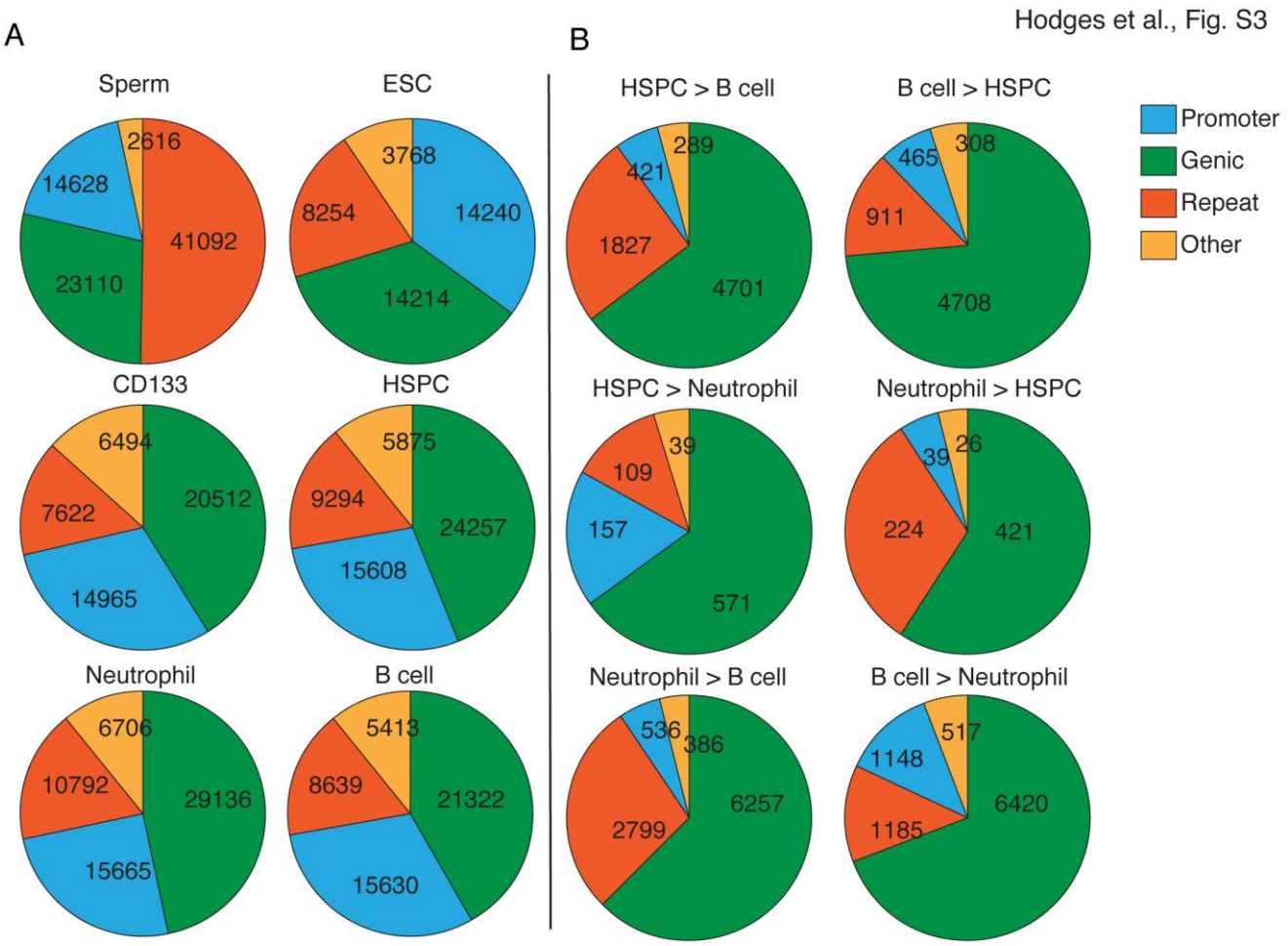
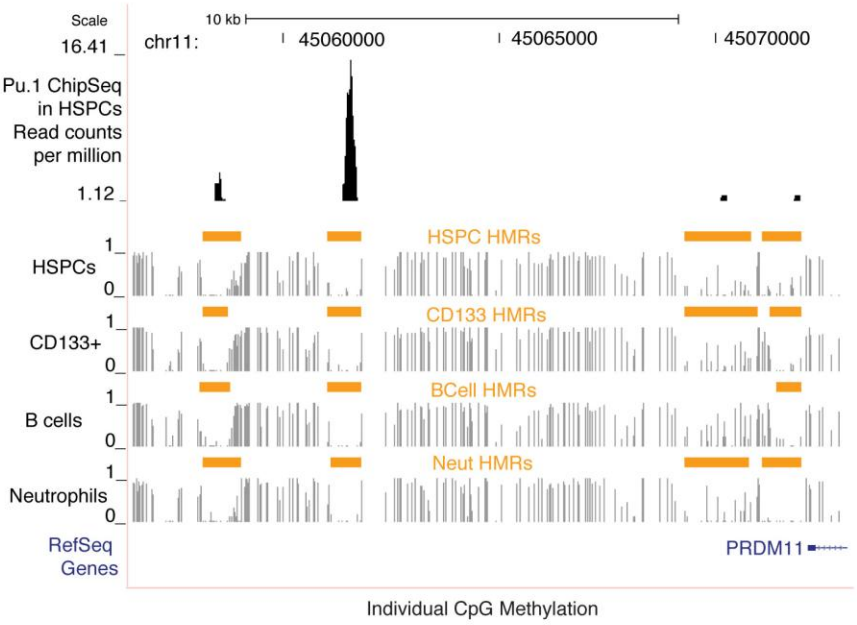


Figure S3. Distribution of HMRs and DMRs According to Genomic Annotations, Related to Figures 2, 4, and 5

The categories “promoter,” “genic,” “repeat” and “other” are exclusive, so first an HMR **(A)** or DMR **(B)** is checked for overlap with a promoter, the remainder are checked for overlap with a genic region, then the remainder are checked for overlap with annotated repeats (any class), and the “other” category is all those that remain.

A


Scale
10 kb
chr11:
45060000
45065000
45070000
16.41
Pu.1 ChipSeq
in HSPCs
Read counts
per million
1.12
HSPC HMRs
HSPCs
CD133 HMRs
CD133+
BCell HMRs
B cells
Neut HMRs
Neutrophils
RefSeq
Genes
PRDM11
Individual CpG Methylation

B

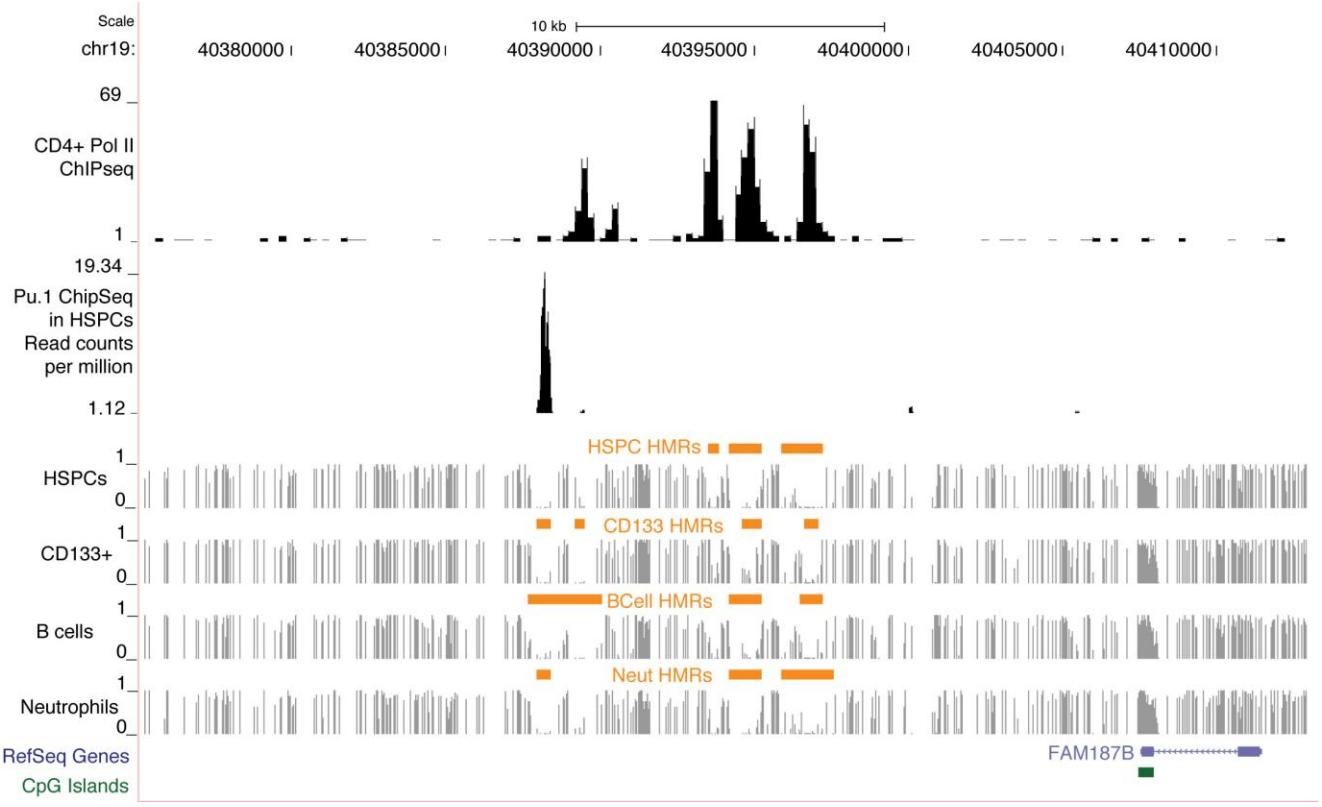

Scale
10 kb
chr19:
40380000
40385000
40390000
40395000
40400000
40405000
40410000
69
CD4+ Pol II
ChIPseq
1
19.34
Pu.1 ChipSeq
in HSPCs
Read counts
per million
1.12
HSPC HMRs
HSPCs
CD133 HMRs
CD133+
BCell HMRs
B cells
Neut HMRs
Neutrophils
RefSeq Genes
FAM187B
CpG Islands

Figure S4. PU.1 and RNA Polymerase II Enrichment in Intergenic HMRs, Related to Figure 4

Sequencing tracks of two loci **(A, B)** with ChIP-seq peaks derived from HSPCs enriched for PU.1 transcription factor (Novershtern et al., Cell. 2011 Jan 21;144(2):296-309) or RNA pol II overlapping intergenic HMRs. Peaks are displayed as read counts per million.

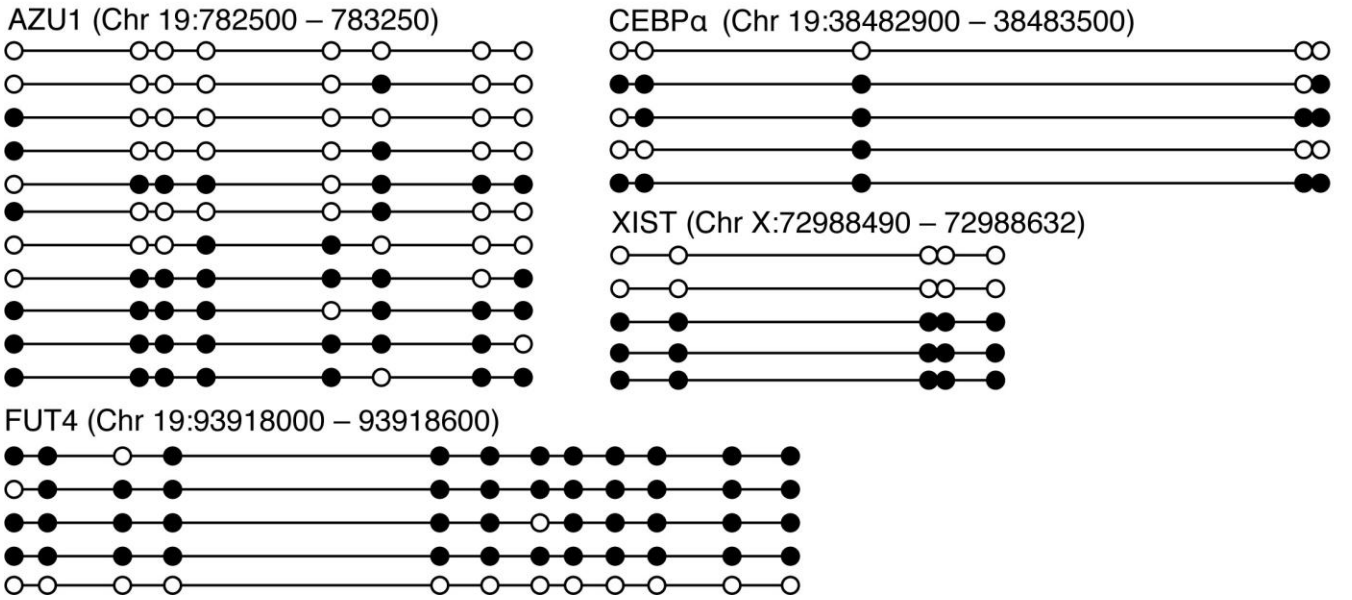


Figure S5. Bisulfite Sequencing of Clones Using Sanger Sequencing, Related to Figure 5

Lollipop diagrams show individual clones derived from HSCs across three myeloid specific genes and an allelicly methylated gene (see also Fig. 5).

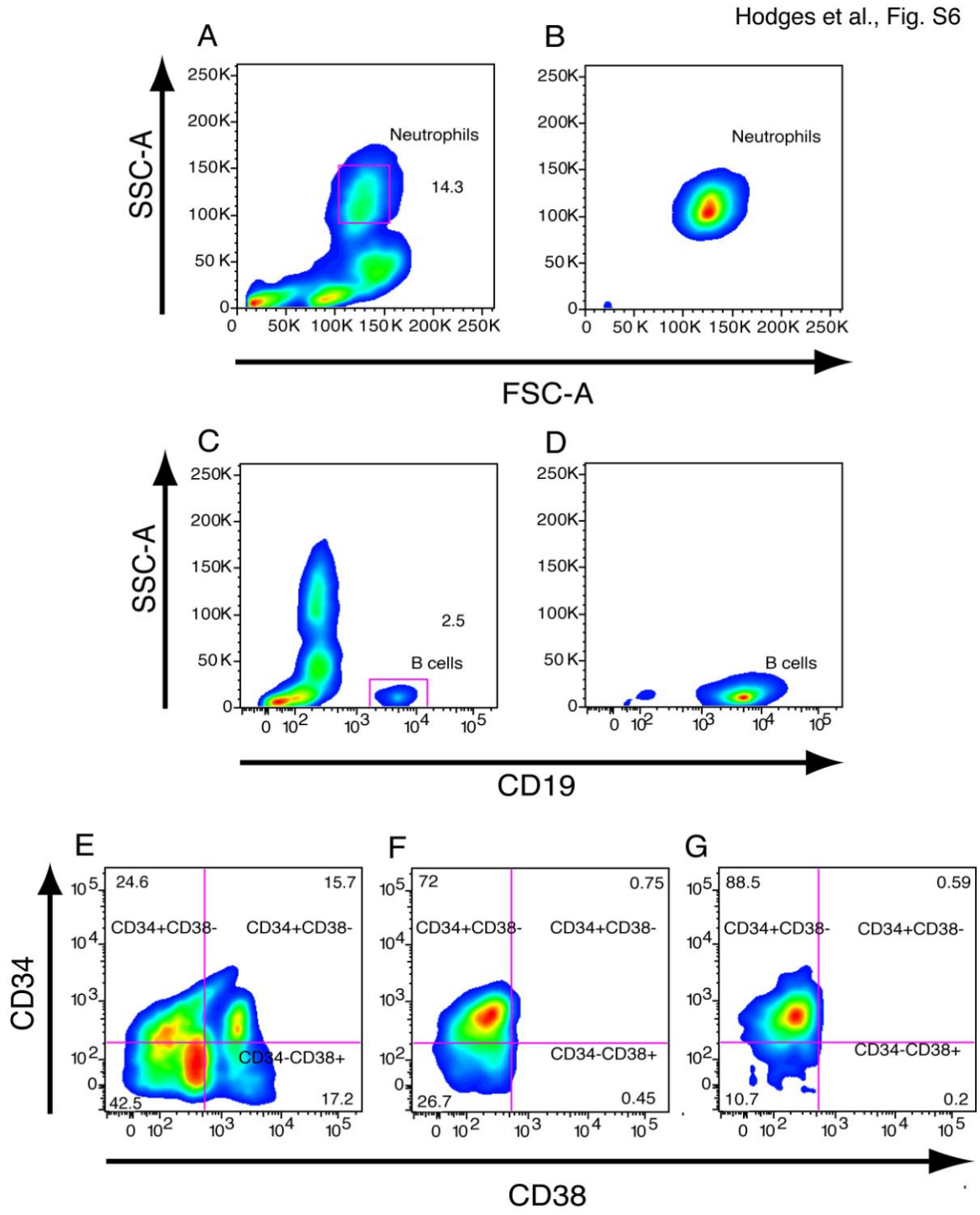


Figure S6. FACS profiles of purified blood cells, related to Figure 5. Peripheral blood mononuclear cells were purified according to the cell surface markers conjugated to the specified fluorophores. Displayed here are Neutrophils before **(A)** and after **(B)** sorting, B cells before **(C)** and **(D)** after sorting. Lineage depleted HSPCs **(E)** underwent two rounds of post-sorting to improve purity levels **(F, G)**.

Table S3. Primers Uses for Bisulfite PCR Cloning, Related to Figure 5

Chr19_CEBP Alpha

Forward – GGA AAG GGA GTT TTA GAT TTT TTT T
Reverse – CTA ACC TCT ATA CCC CAA CAA TAC CT

ChrX_XIST2

Forward – AAA AAG TGT AGA TAT TTT AGA GAG TGT AAT
Reverse – ACT TTA ATT TTT ATT TTT CTA ACC CAT C

Chr19_AZUI

Forward – GGG TTT GTG ATT TTT TAT GGA GTT
Reverse – CTT TAT TAC AAC CAA AAC CCC TCT A

Chr11_FUT4

Forward – GTG GTA TGG GTG GTG AGT TAT T
Reverse – CCA CTA TAT ACA AAA ACC CAA TTT C