Supplemental Tables

Please see associated spreadsheets

Supplementary Table 1: Differentially expressed genes with HSC aging and sequencing statistics. (Related to Figure 1)

Supplementary Table 2: GSEA analysis of differentially expressed genes. (Related to Figure 1)

Supplementary Table 3: Analysis of TGF-β associated gene expression changes in aging HSCs. (Related to Figure 1)

Supplementary Table 4: Isoform switch and transposon elements (TE) expression alterations with HSC aging. (Related to Figure 1)

Supplementary Table 5: Histone marks changed with age. (Related to Figure 2 and Figure 3)

Supplementary Table 6: DMCs in different genomic and epigenomic features. (Related to Figure 4 and Figure 5)

Supplementary Table 7: Analysis of aging associated differential methylated regions (DMRs). (Related to Figure 5 and figure 6)

EXTENDED EXPERIMENTAL PROCEDURES

Hematopoietic Stem Cell Purification and Flow Cytometry

HSCs throughout this study were purified as SP-KSL-CD150$^+$ (see methods) as HSCs defined in this manner are found in both young and aged mice, have high phenotypic homogeneity, largely overlap with HSCs purified via alternative strategies, and exhibit high functional activity when tested by single-cell, or low cell number, transplantation (Camargo et al., 2006; Challen et al., 2010; Chambers et al., 2007b; Ergen et al., 2012; Mayle et al., 2012). Whole bone marrow cells were isolated from femurs, tibias, pelvis and humerus. SP staining was performed with Hoechst 33342 (Sigma) as previously described (Goodell et al., 1996). Briefly, whole bone marrow cells were resuspended in staining media at 10^6 cells/mL and incubated with 5 mg/ml Hoechst 33342 for 90 minutes at 37^{o}C. For antibody staining, cells were suspended at a concentration of 10^8 cells/ml and incubated in 4°C for 15 minutes with the desired antibodies. Magnetic enrichment was performed with c-Kit-biotin antibody (eBioscience, San Diego, CA) and anti-biotin microbeads (Miltenyi Biotec, Auburn, CA) or anti-mouse CD117 microbeads (Miltenyi Biotec, Germany) on an AutoMACS (Miltenyi Biotec, Germany). Post-enrichment, the positive cell fraction was labeled with antibodies to identify HSCs (SP$^+$ Lineage (CD3, CD4, CD8, B220, Gr1, Mac1 and T119)$^-$ Sca-1$^+$ c-Kit$^+$ CD150$^+$), B cells (B220$^+$) and Gr cells (Gr1$^+$). All antibodies were obtained from BD Biosciences (San Jose, CA) or eBioscience (San Diego, CA) and used at 1:100 dilutions. Cell sorting was performed on a MoFlo cell sorter (Dako North America, Carpinteria, CA) or Aria II (BD Biosciences, San Jose, CA) and analysis performed on a LSRII (BD Biosciences, San Jose, CA).

RNA-sequencing

Batches of approximately 70,000 HSCs, 1 million B cell and Gr cells were FACS sorted. RNA was isolated with the RNeasy Micro kit (Qiagen, Valencia, CA), including the DNase I (Qiagen) on-column digestion. Paired end libraries were generated by using Illumina TruSeq RNA sample preparation kit. Illumina HiSeq was used for sequencing with a paired-end sequencing length of 100bp.

ChIP-sequencing (ChIP-seq)

Chromatin Immunoprecipitation (ChIP) was performed with 50,000 to 100,000 HSCs, B-cells, and Granulocytes according to standard protocols adapted for small cell numbers; see Extended methods_ENREF_4. ChIPed DNA was successfully made into sequencing libraries using the ThruPLEX-FD preparation kit without extra amplification (Rubicon, Ann Arbor, MI). Sequencing was performed according to the manufacturer's protocol on a HiSeq 2000 (Illumina). Sequenced reads were mapped to the mm9 mouse genome and peaks were identified by model-based analysis of ChIP-seq data (MACS).

Whole-genome bisulfite sequencing (WGBS)

For WGBS library construction, 300ng genomic DNA was isolated from HSCs and fragmented using a Covaris sonication system (Covaris S2). Following fragmentation, libraries were constructed using the Illumina TruSeq DNA sample preparation kit. After ligation, libraries were bisulfite-treated using the EpiTect Bisulfite Kit (Qiagen, Valencia, CA). Ligation efficiency tested by PCR using TrueSeq primers and Pfu TurboCx hotstart DNA polymerase (Stratagene). After determining the optimized PCR cycle number for each samples, a large scale PCR reaction (100 µl) was performed as described previously (Gu et al., 2011). PCR products were sequenced with Illumina HiSeq sequencing systems.

Quantitative Real-Time PCR

RNA was isolated using the RNeasy Micro kit (Qiagen). First-strand cDNA was synthesized by SuperScript II reverse transcriptase (Invitrogen). cDNA input was standardized and RT-PCRs were performed with Taqman master Mix (Applied Biosystems, Carlsbad, CA), 18s-rRNA probe (VIC-MGB; Applied Biosystems), and a gene-specific probe (FAM-MGB; Applied Biosystems) for 40 cycles with an AbiPrism 7900HT (Applied Biosystems). Samples were normalized to 18S and fold-change determined by the delta Ct method.

Chromatin Immunoprecipitation (ChIP)

HSCs (50,000~100,000), B-cells, and Granulocytes were sorted and crosslinked with 1% formaldehyde at room temperature (RT) for 10 min, and the reaction was stopped by adding Glycine to a final concentration 0.125M and incubated at room temperature for 5 min. Then the cells were washed once with ice cold PBS containing protease inhibitor cocktail (PIC; Roche) and the pellet was stored at −80°C. The pellet was thawed on ice and lysed in 50 µl lysis buffer (10 mM Tris pH 7.5, 1mM EDTA, 1% SDS), then diluted with 150 µl of PBS/PIC, and sonicated to 200-500 bp fragments (Bioruptor, Diagenode). The sonicated material was centrifuged at 4°C for 5 min at 13,000 g to remove precipitated SDS. An aliquot of the supernatant (typically 180 µl) was then transferred to a new 0.5 ml collection tube, and 180 µl of 2X RIPA buffer (20 mM Tris pH 7.5, 2 mM EDTA, 2%Triton X-100, 0.2% SDS, 0.2% sodium deoxycholate, 200 mM NaCl/PIC) was added. A 1 /10 volume (36 µl) was removed for input control. ChIP-qualified antibodies (0.1 µg H3K4me3 Millipore 07-473, 0.3 µg H3K27me3 Millipore 07-449) were added and the mixture was incubated at 4°C overnight. Following this, 10 µl of protein A magnetic beads (Dynal, Invitrogen) previously washed in RIPA buffer were added and the mixture was incubated for an additional 2 hours at 4°C. The bead:protein complexes were washed three times with RIPA buffer and twice with TE (10 mM Tris pH 8.0/1 mM EDTA) buffer using centrifugation to precipitate the complexes. Following transfer into new 1.5 ml collection tube, genomic DNA was released during 2 hours at 68 °C in 100 µl Complete Elution Buffer (20 mM Tris pH 7.5, 5 mM EDTA, 50 mM NaCl, 1% SDS, 50 µg/ml proteinase K), and combined with a second treatment of 100 µl Elution Buffer (20 mM Tris pH 7.5, 5 mM EDTA, 50 mM NaCl) for 10 min at 68 °C. ChIPed DNA was purified by MinElute Purification Kit (Qiagen) and eluted in 12 µl elution buffer.

High-Performance Liquid Chromatography – Mass Spectrometry

Two µg of genomic DNA was digested with a cocktail of nuclease enzymes using a commercial kit following the manufacturers' protocol (DNA Degradase Plus, Zymo Research; 2.5 µl 10X DNA Degradase Reaction buffer, 1 µl DNA Degradase Plus and water to make a total reaction volume of 25 µl). After incubation (37 ºC, >1 hr) aqueous formic acid was added (25 µl, 0.1% v/v) to yield a final concentration of 40 ng of digested DNA/µl.

Three 897bp DNA standards, each homogenous for either unmodified 2'-deoxycytidine (dC), 5-methyl-2'deoxycytidine (5mdC), or 5-hydroxymethyl-2'-deoxycytidine (5hmdC), were purchased (Zymo, Irvine, CA), and used to generate a calibration curve. The standards had been prepared by PCR using the appropriate nucleotides and were spin-column purified by the manufacturer to obtain 50 ng/uL aqueous Tris buffered solutions. By MRM criteria these standards were all more than 99.6% pure. With each batch of experimental samples a series of standard samples was simultaneously prepared using the DNA standards. The standard samples contained increasing amounts of 5mdC and 5hmdC in the presence of the same amount of dC (0, 0.1, 1, 5 and 10% for 5mdC and 0, 0.1, 0.5, 1, and 2% for 5hmdC).

The MRM quantitation method was slightly modified from that described previously (Le et al., 2011). DNA hydrolysis samples were injected onto a reverse phase UPLC column (Eclipse C18 2.1 x 50 mm, 1.8 µm particle size, Agilent) equilibrated with buffer A (0.1% aqueous formic acid) and eluted (200 µL/min) with an increasing concentration of buffer B (methanol: min/%B; 0/0, 2/0, 4/5, 6/5, 8/0, 10/0). The injection volume for each sample is adjusted such that the dC peak area was at least 1 million area counts (Agilent MassHunter Quantitative Analysis, version B.04.00), which is equivalent to 100 ng of digested DNA. The effluent from the column was directed to an electrospray ion source (Agilent Jet Stream) connected to a triple quadrupole mass spectrometer (Agilent 6460 QQQ) operating in the positive ion multiple reaction monitoring mode using previously optimized conditions, and the intensity of specific MH$^+$→fragment ion transitions were recorded (5mdC m/z 242.1→126.1, 5hmdC 258.1→142.1 and dC m/z 228.1→112.1).

Calibration curves were constructed for 5mdC and 5hmdC from the data obtained from the standard samples (measured 5mdC or 5hmdC peak area/total cytosine pool plotted against actual percentage of either 5mdC or 5hmdC in the samples). The measured percentage of 5mdC and 5hmdC in each experimental sample was then converted to actual percentage 5mdC and 5hmdC by interpolation from the calibration curves. This provided a correction for any differences that might exist in the molar MRM responses of the various nucleosides.

miRNA cloning and retrovirus transduction

We used software Block-iT RNAi Designer (Invitrogen) to design miRNAs targeting *Slc22a3*. The stem-loop hairpin produces a miRNA that 100% matches to the gene of interest and cleaves the target mRNA. Oligos targeting each novel transcript were successfully cloned by BLOCK-iT PolII miR RNAi Expression Vector Kit (Invitrogen). Oligos targeting lacz were provided by the kit and used as control in all experiments. Briefly, the synthetic double-stranded oligos were cloned into the vector, pcDNA 6.2-GW/EmGFP-miR. The stem-loop hairpin with GFP tag was then incorporated into the pDonor vector using BP clonase enzyme mixture (Invitrogen). The oligos were further recombined into the retroviral MSCV-RFB vector (containing attR recombination sites) using LR clonase enzyme mixture (Invitrogen). Viruses were packaged by cotransfection with pCL-Eco into 293T cells. Viral supernatants were collected 48-hours post-transfection and viral titers determined using 3T3 cells.

For retroviral transduction of hematopoietic progenitors, donor mice were treated with 5-fluorouracil (150mg/kg; American Pharmaceutical Partners, Schaumburg, IL) six days prior to bone marrow harvest. Whole bone marrow was enriched for Sca-1$^+$ cells using magnetic enrichment (AutoMACS; Miltenyi Biotec) and adjusted to a concentration of 5 x 10^5 cells/ml in transduction medium, containing Stempro 34 (Gibco, Carlsbad, CA), nutrient supplement, penicillin/streptomycin, L-glutamine (2mM), mSCF (10ng/ml; R&D Systems Minneapolis, MN), mTPO (100ng/ml; R&D Systems). The suspension was spin-infected at 250 x g at room temperature for 2 hours in the presence of polybrene (4 µg/ml). For in vivo transplantation, cells were incubated for a further 1 hour at 37^{o}C. For *in vitro* assays, transduced cells were cultured in fresh transduction medium for a further two days.

miRNA Oligos: (target sequence underlined):

Lacz-F:
TGCTGAAATCGCTGATTTGTGTAGTCGTTTTGGCCACTGACTGACGACTACACATCAGCGATTTCAGGACACAAGGCC

Lacz-R:

5'-
CCTGAAATCGCTGATGTGTAGTCAGTCAGTCAGTGGCCAAAACGACTACACAAATCAGCGA
TTTC-3'

Slc22a3-F
5'-
TGCTGAAATCTTTACGGTTCCTTGGAGTTTTGGCCACTGACTGACTCCAAGGACGTAAAGA
TTT -3'

Slc22a3-R
5'-
CCTGAAATCTTTACGTCCTTGGAGTCAGTCAGTGGCCAAAACTCCAAGGAACCGTAAAGAT
TTC -3'

RT-PCR primers:

Slc22a3-RT-F: AATATCCTGTTTCGGCGTTG
Slc22a3-RT-R: TCACGAAGCAAGTCATCCAG

In vivo Transplantation

All mice were C57Bl/6 background distinguished by CD45.1 or CD45.2 alleles. For bone marrow transplantation, recipient C57Bl/6 CD45.1 mice were transplanted by retro-orbital injection following a split dose of 10.5 Gy of lethal irradiation. 50,000 Sca-1^{+} (CD45.2) donor cells were injected to the recipient mice.

Peripheral Blood Analysis

For peripheral blood analysis by flow cytometry, mice were bled retro-orbitally, the red blood cells were lysed, and each sample was incubated with the following antibodies on ice for 20 min: CD45.1-FITC, CD45.2-APC, CD4-Pacific Blue, CD8-Pacific Blue, B220-Pacific Blue, B220-PeCy7, Mac1-PeCy7, and Gr-1-PeCy7 as previously described (Mayle et al., 2013). Cells were then spun down and re-suspended in a propidium iodide solution, and analysis was accomplished on live cells with an LSRII (Becton Dickinson).

Gene expression analysis of RNA-Seq data

The reads are all 100bp long and paired-ended. The last 20 bases are trimmed due to average low quality. The alignment was performed by RUM (Grant et al., 2011), which first tries to map reads to genome and transcriptome by Bowtie, and then the reads unmapped to genome are handed to Blat for additional mapping. The information from the three mappings is merged into one mapping. The multiply mapped reads are then discarded. The gene annotations used for transcriptome alignment include refSeq gene model, UCSC knownGene model and ensemble gene model. The gene expression, FPKM value, is calculated by counting the fragments matching the exon information for each gene. Differential expression was performed using DESeq (Anders and Huber, 2010). By using p-value cutoffs of 0.05, 1337 up-regulated and 1297 down-regulated genes are discovered. We used DAVID to examine these differentially expressed genes for functional enrichment in GO terms, KEGG Pathways, and SP_PIR_KEYWORDS. The unbiased Gene Set Enrichment Analysis was performed using GSAA-Seq (http://gsaa.unc.edu), which ranks all genes by DESeq differential test p-values and examines enrichment of all gene sets in the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005), and several manually created hematopoiesis fingerprint gene sets (Chambers et al., 2007a).

IPA analysis of differentially expressed genes

The TGF-β signaling reduction was identified by IPA (Ingenuity Systems, www.ingenuity.com). Of 1238 genes meeting a threshold of fold-change > ±1.5 and multiple testing corrected p-value of < 0.001, 1121 were characterized in the Ingenuity database and thus were included in the analysis. The upstream regulator analysis and mechanistic network discovery algorithms are based on prior knowledge of expected causal effects between transcriptional regulators and their target genes documented from the literature compiled in the Ingenuity database. The database contains a causal network with ~ 39,000 nodes (genes, miRNA, chemicals) and ~ 116,000 edges. In IPA, edges represent experimentally observed cause-effect relationships (direct or indirect) and binding events. Typically edges are associated with a direction of regulation, either "activating" or "inhibiting". An upstream regulator (broadly defined as any molecule which can influence gene expression) is connected to a gene dataset through expression edges (Kramer et al., 2012).

For each potential regulator, two statistical measures, an overlap p-value and an activation z-score are calculated. The overlap p-value measures the significance of overlap between the known targets of each regulator and a dataset (Fisher's Exact Test, right-tailed). The activation z-score is an independent metric to call upstream regulators by inferring their activation state. The basis for inference is expression edges in Ingenuity's causal network. Given the observed differential expression direction of a gene in the dataset, the activation state of an upstream regulator is determined by the regulation direction associated with the relationship from the regulator to the gene compared to a model that assigns random regulation directions.

We next constructed a mechanistic network of regulators cooperating with TGFB1. We identified all regulators (overlap p-value < 1.0E-8, z-score > 1.5) that are connected downstream of TGFB1 through one edge (edge p-value < 5.0E-8) where the edge p-value is based on the overlap between the corresponding regulated genes in the dataset (Fisher's Exact Test, right-tailed). The network represents the union of all paths connecting regulators to the dataset.

Repeat analysis of RNA-Seq data

Repeat elements annotation was downloaded from UCSC RepeatMasker Table (Dec 2011). When reads were aligned to those repeat regions, many were mapped to multiple locations. A statistical framework RSEM (RNA-Seq by Expectation Maximization) (Li et al., 2010) was used to assign the reads probabilistically to address this mapping uncertainty. RSEM was originally developed to handle the multiple mapped reads in gene/isoform estimation, which was similar to our repeat elements expression quantification. During calculation, RSEM would maintain a list of counts of number of reads assigned to each repeat element. On each iteration of the algorithm, the probability of a read assigned to each of its possible positions was calculated based on the counts from the previous iteration. After getting the number of reads in each repeat, we employed edgeR to call differentially expressed repeat elements with false discovery rate 0.05.

Alternative splicing of RNA-Seq data

PSI (percentage spliced in) was used to denote the fraction of mRNAs that retained the alternatively spliced cassette exon (Katz et al., 2010). Here, SpliceTrap (Wu et al., 2011) was employed to estimate PSI from paired-end RNA-Seq data. If the difference of PSIs of two conditions was no less than 0.1, this event was reported as differential alternative splicing.

Alternative pre-mRNA of RNA-Seq data

A number of paired-end fragments (X) can be identified confidently from pre-mRNA if the fragment does not splice and spans exon intron junction. Similarly a number of fragments (Y) can be identified confidently from mRNA if the fragment splices across the exon/intron junction. The pre-mRNA abundance is X/(X+Y). We have ignored the fragments that cannot be confidently classified, such as the fragment contained in intron. The abundance difference in young and old are reported if the difference is larger than 0.2 and the p-value from Fisher's Exact Test is significant.

Peak calling of histone modification data

The young and old samples were sequenced multiple times. The reads are mapped to mouse genome mm9 using SOAP2 (Li et al., 2009) by allowing at most 2 mismatches for 50bp long short reads and at most 4 mismatches for 100bp long short reads. Only uniquely mapped reads were retained. To remove duplicate reads resulting from the PCR amplification, at most 2 duplicate reads were allowed for each biological replicate. The number 2 is based on Poisson P-value cutoff of $1x10^{-5}$ determined by the total number of reads with respect to the theoretical mean coverage across the genome. The uniquely mapped and duplicate removed reads from each biological replicate are fed as a treatment file into the MACS program (Zhang et al., 2008), to find the enriched regions, or "peaks". Peaks are regions with enrichment of treatment reads compared to control reads, which are just sonicated wild-type or knock-out sample DNA fragments without ChIP pull down. The p-value cutoff for MACS is E-8. Peaks from all biological replicates of a specific sample are merged to form the final set of peaks for this specific sample.

K-means clustering of histone modification data

We then use seqMiner (Ye et al., 2011) to do clustering by pooling all uniquely mapped and duplicate-removed reads. We used the R package 'fpc' (http://cran.r-project.org/web/packages/fpc/index.html) to estimate the optimal number of clusters for the k-means clustering. This estimation gives either 3 or 4 depending on what the histone modification is. The genomic regions of interest are from -5kb of TSS to +5kb of TSS or from TSS to TTS of Refseq genes.

Quantitative analysis of differential histone methylation regions

The quantitative comparison analysis was done by comparing the peak signals between the two ChIP samples (young and old). To do so, we merged the young and old peaks, counted the reads on each of the merged regions, and performed the Poisson test as in the MACS software. For each test, the two numbers under test are the numbers of reads per million total reads for young and old samples for each specified region. Suppose the number for young sample is the mean value of the Poisson distribution, the p-value is calculated for the number of the old sample. We then perform the test again by swapping the two numbers in the test formula, i.e., regarding the numbers for the old sample as mean value while the number for the young sample as the test number. A region with p-value of less than E-8 in one of the two tests was considered as a region with a significant difference between young and old. In addition to the quantitative analysis on peaks, the comparison on promoters of Refseq genes was also performed in order to detect the significantly changed promoters. The procedure on the promoters is same for that on merged peaks.

Analysis of Whole Genome Bisulfite Data

The WGBS data analyses were based on a newly developed program MOABS: MOdel based Analysis of Bisulfite Sequencing (Sun et al. http://code.google.com/p/moabs/, manuscript in preparation). We have used four modules of MOABS, mMap, mCall, mOne and mComp from this software. MOABS seamlessly integrates alignment, methylation ratio, and identification of hypomethylation for one sample and differential methylation for multiple samples, and other downstream analysis.

Reads Mapping

mMap module used the default mapping program BSMAP(Xi and Li, 2009) to align the paired-end sequences onto mouse genome mm9. The adaptor and low quality end of read is automatically trimmed by BSMAP. For each read, the mapping location was determined to be the location with the fewest mismatches. If a read can be mapped to multiple locations with same fewest mismatches, this read is determined as a multiply mapped read and its mapping location is randomly selected from all the locations with fewest mismatches.

Quality control and methylation ratio calling

For reads from each library, we allow at most 2 reads mapped to exactly same location and remove any extra reads which are regarded as PCR amplification products from same DNA fragment. During the preparation of the bisulfite treated library, unmethylated cytosines residues were added to do the end repair. This end repair procedure may introduce artifact if the repaired bases contain methylated cytosine. We modeled the length of overhang sizes of read fragments and determined that trimming 3 bases from the repaired end would eliminate nearly all of potential artifact events. We also enabled the process PEOverlap option of mcall module so that the overlapping segment of two read mates is only processed once to prevent over-counting the same information. Methylation ratio for each individual CpG is measured as total number of unconverted CpGs divided by the total number unconverted CpGs and TpGs at this specific location, counting both strands. Bisulfite conversion rate is measured as the total number of CpH to TpH bisulfite transitions, divided by the total number of CpHs in the mapped reads.

Differentially methylated regions (DMRs)

Here, mSuite used a first order Hidden Markov Model to find differentially methylated regions. The state of i^{th} cytosine in the genome is denoted as S_i where S_i can take 3 hidden states for a two-sample comparison:

S_0: hypo-methylation state if $p_2 - p_1 < -v_0$;

S_1: no difference state if $|p_2 - p_1| < v_0$;

S_2: hyper-methylation state if $p_2 - p_1 > v_0$;

where v_0 is a preset parameter and marks the characteristic threshold of difference for underlying dataset.

We model the neighbor correlation by first order Markov chain

$$\Pr(S_i) = \Pr(S_i|S_{i-1}),$$

which means that the state of site i is directly influenced by previous site.

Each observation for each site is a combination of 4 numbers from 2 samples: $x = (n_1, k_1, n_2, k_2)$. In this problem, we are given the observation sequence from all sites, we want to find the HMM model that maximizes the probability of observation sequence. The HMM is characterized by initial state π_0, transition probability matrix $A = Pr\left(S_i \mid S_{i-1}\right)$ and emission probability matrix $B = Pr\left(x_i \mid S_i\right)$.

The initial state π_0 can just takes value S_1, though its value does not matter since there are millions of CpGs in the genome.

By assuming a site is in one of the three states, the emission probability for the i^{th} site to observe $x = (n_1, k_1, n_2, k_2)$ when the state of the site is S_i, can be derived as

$$\Pr(n_1, k_1, n_2, k_2 \mid s_i) = \frac{\iint_{s_i} dp_2\, dp_1 f(k_1; n_1, p_1) f(k_2; n_2, p_2)}{\int_0^1 f(k_1; n_1, p_1)\, dp_1 \int_0^1 f(k_2; n_2, p_2)\, dp_2}$$

Since there are millions of sites and there is a high chance of repeated observations, mSuite uses a lookup table to avoid repeated computation of numerical integrations.

The state transition probability matrix can be trained using the forward-backward algorithm. In the training process, the initial state, and the emission probability matrix are fixed while the state transition probability is the only model variable. Since the training is computationally intensive, mSuite may choose only a subset of all cytosine sites in the genome, like 1st one million sites or all sites in chromosome 19 or locus provided by users. After the change of likelihood of the model is smaller than a given threshold or max number of iterations is reached, the optimal hidden state for each site is obtained. Consecutive sites with S_0 (or S_1) states are merged as hypo-DMR (or hyper-DMR).

DMRs functional enrichment and overlap with TF binding sites

A DMR is assigned to a gene if the DMR is located within 3kb of the gene body. Then we use DAVID to perform the functional enrichment analysis for the hypo-DMR or hyper-DMR marked genes respectively. We also obtained hundreds of bed files with each representing binding sites of transcription factor in a blood cell (Hannah et al., 2011). The overlap between each transcription factor binding sites (TFBS) and DMRs are calculated. Dividing the genome into TFBS and non-TFBS, and assuming the NULL distribution of DMRs is uniform across the genome, we calculated the p-value from the hyper-geometric distribution function.

REFERENCES FOR SUPPLEMENTARY MATERIALS

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Camargo, F.D., Chambers, S.M., Drew, E., McNagny, K.M., and Goodell, M.A. (2006). Hematopoietic stem cells do not engraft with absolute efficiencies. Blood *107*, 501-507.

Challen, G.A., Boles, N.C., Chambers, S.M., and Goodell, M.A. (2010). Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. Cell Stem Cell *6*, 265-278.

Chambers, S.M., Boles, N.C., Lin, K.Y., Tierney, M.P., Bowman, T.V., Bradfute, S.B., Chen, A.J., Merchant, A.A., Sirin, O., Weksberg, D.C.*, et al.* (2007a). Hematopoietic fingerprints: an expression database of stem cells and their progeny. Cell Stem Cell *1*, 578-591.

Chambers, S.M., Shaw, C.A., Gatza, C., Fisk, C.J., Donehower, L.A., and Goodell, M.A. (2007b). Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. PLoS Biol *5*, e201.

Ergen, A.V., Boles, N.C., and Goodell, M.A. (2012). Rantes/Ccl5 influences hematopoietic stem cell subtypes and causes myeloid skewing. Blood *119*, 2500-2509.

Goodell, M.A., Brose, K., Paradis, G., Conner, A.S., and Mulligan, R.C. (1996). Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. J Exp Med *183*, 1797-1806.

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., and Pierce, E.A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics *27*, 2518-2528.

Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A., and Meissner, A. (2011). Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc *6*, 468-481.

Hannah, R., Joshi, A., Wilson, N.K., Kinston, S., and Gottgens, B. (2011). A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. Experimental hematology *39*, 531-541.

Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods *7*, 1009-1015.

Kramer, A., Tugendreich, S., and Green, J. (2012). Mechanistic networks: explaining gene expression data using literature-based molecular interactions. In 5th Annual RECOMB Conference on Regulatory Systems Genomics, with DREAM Challenges An Official Conference of the International Society for Computational Biology (San Francisco, CA)

Le, T., Kim, K.P., Fan, G., and Faull, K.F. (2011). A sensitive mass spectrometry method for simultaneous quantification of DNA methylation and hydroxymethylation levels in biological samples. Anal Biochem *412*, 203-209.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493-500.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Mayle, A., Luo, M., Jeong, M., and Goodell, M.A. (2012). Flow cytometry analysis of murine hematopoietic stem cells. Cytometry Part A : the journal of the International Society for Analytical Cytology.

Mayle, A., Luo, M., Jeong, M., and Goodell, M.A. (2013). Flow cytometry analysis of murine hematopoietic stem cells. Cytometry A *83*, 27-37.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S*., et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550.

Wilson, N.K., Foster, S.D., Wang, X., Knezevic, K., Schutte, J., Kaimakis, P., Chilarska, P.M., Kinston, S., Ouwehand, W.H., Dzierzak, E*., et al.* (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell *7*, 532-544.

Wu, J., Akerman, M., Sun, S., McCombie, W.R., Krainer, A.R., and Zhang, M.Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics *27*, 3010-3016.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC bioinformatics *10*, 232.

Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res *39*, e35.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W*., et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.