doi:10.1038/nature14465

Methods

Additional files referred to throughout these methods can be found here:

http://neomorph.salk.edu/SDEC_tissue_methylomes/processed_data/code_data.tar.gz

Tissue Collection

Adrenal, adipose, thymus, esophagus, vascular, bladder, pancreas, liver, stomach, lung, heart, skeletal muscle, ovary, small bowel, colon, and spleen tissues were obtained from deceased donors at the time of organ procurement at Mid-American Transplant Services (St. Louis, USA) after research consent from family was obtained. Samples were flash frozen with liquid nitrogen. From the following tissues, the luminal epithelial lining was dissected free and flash frozen for this study: esophagus, bladder, stomach, small bowel and colon. For tissue from the aorta, the endothelial layer was dissected free and flash frozen.

Genomic DNA Sequencing Library Construction

Two µg of genomic DNA was extracted from ground, frozen tissue using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA). The DNA was fragmented with a Covaris S2 (Covaris, Woburn, MA) to 300-400 bp, followed by library preparation using the TruSeq DNA Sample Prep kit (Illumina, San Diego, CA) as per manufacturer's instructions. The library was run on a 2% agarose gel and gel size selected to 400-500bp using the MinElute Gel Extraction kit (Qiagen).

RNA-seq Library Construction

Total RNA from tissues and primary cells was extracted using the RNeasy Lipid Tissue Mini Kit according to protocol (QIAGEN). The mRNA libraries were constructed using the TruSeq RNA Sample Prep Kit V2 (Illumina, San Diego, CA) with 4 µg total RNA, according to manufacturer's instructions with modifications to confer strand specificity. The RNA was incubated in the Elute, Prime, Fragment Mix at 94˚C for 4 min. After first strand synthesis, the product was purified using RNAClean XP beads (Beckman, Brea, CA) as per manufacturer's instructions and eluted in 18 µL nuclease free water. Second strand synthesis was performed by adding the RNAClean

XP purified product to 2.5 µL 10x NEB Buffer 2 (New England Biolabs, Ipswich, MA), 2 µL dUTP mix (10mM dATPs, 10mM dGTPs, 10mM dCTPs, and 20mM dUTPs), 0.5 µL RNAse H (2 U/µL), 1 µL DNA Polymerase I (E. coli) (New England Biolabs), and 1 µL DTT (100 mM). The 25 µL mixture was incubated at 16˚C for 2.5 hours. The purified ligation products were incubated with 2 µL Uracil DNA Glycosylase (Fermentas) before PCR amplification. The completed library was then gel size selected to approximately 350-450 bp using the QIAquick Gel Extraction Kit (QIAGEN). RNA-seq libraries were sequenced using the Illumina HiSeq 2000 (Illumina) instrument as per manufacturer's instructions. Sequencing of libraries was performed up to 2 × 101 cycles. Image analysis and base calling were performed with the standard Illumina pipeline version RTA 2.8.0

MethylC-seq Library Construction

Genomic DNA was extracted from ground, frozen tissue using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). Two µg of genomic DNA was spiked with 10 ng unmethylated cl857 Sam7 Lambda DNA (Promega, Madison, WI). The DNA was fragmented with a Covaris S2 (Covaris, Woburn, MA) to 150-200 bp, followed by end repair and addition of a 3' A base. Cytosine-methylated adapters provided by Illumina (Illumina, San Diego, CA) were ligated to the sonicated DNA at 16˚C for 16 hours with T4 DNA ligase (New England Biolabs). Adapter-ligated DNA was isolated by two rounds of purification with AMPure XP beads (Beckman Coulter Genomics, Danvers, MA). Adapter-ligated DNA (≤450 ng) was subjected to sodium bisulfite conversion using the MethylCode kit (Life Technologies, Carlsbad, CA) as per manufacturer's instructions. The bisulfite-converted, adapter-ligated DNA molecules were enriched by 4 cycles of PCR with the following reaction composition: 25 µL of Kapa HiFi Hotstart Uracil+ Readymix (Kapa Biosystems, Woburn, MA) and 5 µl TruSeq PCR Primer Mix (Illumina) (50 µl final). The thermocycling parameters were: 95˚C 2 min, 98˚C 30 sec, then 4 cycles of 98˚C 15 sec, 60˚C 30 sec and 72˚C 4 min, ending with one 72˚C 10 min step. The reaction products were purified using AMPure XP beads. Up to two separate PCR reactions were performed on subsets of the

adapter-ligated, bisulfite-converted DNA, yielding up to two independent libraries from the same biological sample.

SNP Calling

SNPs in each of the four donor genome sequences and the H1 genome were detected as follows. Tissue genome sequence fastq files of four donors were mapped using Bowtie2[1] and its default parameters; whereas, the H1 csfasta files were mapped with Bowtie using these parameters: "-C -k 1 -m 1 --best --strata -e 80". The UnifiedGenotyper module of GenomeAnalyzerTK[2] (GATK) version 2.4-7 was used to detect SNPs. Default parameters were used, with "-dcov 100". The SNPs detected were compared against the dbSNP database (version 137) for classifying known and novel (individual-specific) SNPs. The confidence score threshold for SNP detection was selected as 30. This is the minimum phred-scaled Q-score threshold, provided as a default parameter for high-confidence SNP detection within the GATK package.

SNP-substituted Reference Genomes

We created four modified reference genomes to account for misclassification of CG sites as mCH sites. To that end, we took high-confidence homozygous SNPs and substituted the SNP bases for a particular individual into the hg19 FASTA file.

MethylC-seq Mapping

Sequencing reads were first trimmed for adapter sequence using Cutadapt[3]. All cytosines in the trimmed reads were then computationally converted to thymines and mapped twice, to a converted forward strand reference and to a converted reverse strand reference. A converted reference is created by replacing all cytosines with thymines (forward strand) or all guanines with adenines (reverse strand) in the reference FASTA file. For mapping we used Bowtie[4] with the following options: "-S","-k 1","-m 1","--chunkmbs 3072","--best","--strata","-o 4","-e 80","-l 20", and "-n 0". Reads were mapped to hg19 reference genome. Any read that mapped to multiple locations was removed and one read from each starting location on each strand from

each library was kept (i.e., clonal reads were removed). Note that our pipeline (methylpy) does not currently support paired-end reads. Consequently, for MSC, which only had paired-end reads available, we mapped the first read in each pair to avoid problems in processing overlapping reads.

Methylation Calling

To call methylated sites, we summed the number of reads that supported methylation at a site and the number of reads that did not. We used these counts to perform a binomial test with a probability of success equal to the non-conversion rate, which was determined by computing the fraction of methylated reads in the lambda genome (spiked in during library construction). The false discovery rate (FDR) for a given p-value cutoff was computed by calculating the fraction of sites in the lambda genome that had a p-value less than or equal to the cutoff and then dividing that quantity by the fraction of sites that had a p-value less than or equal to the cutoff across all other chromosomes. Because the p-value distributions for each methylation context are different, this procedure was applied to each three nucleotide context independently (e.g., a p-value cutoff was calculated for CAT cytosines). All methylation data was visualized with the AnnoJ browser[5].

DMR Finding

To find tissue-specific differentially methylated regions (DMRs), we used the method described in Ziller et al.[6] Briefly, a beta-binomial distribution was used to model the methylation level of each single CG site in each of the tissues. Then, differentially methylated sites (DMS) were identified if the methylation levels of certain site were significantly different between tissues (p-value <= 0.01) and the minimum methylation difference was greater than or equal to 0.3. In the next step, DMSs within 500 bp were merged into DMRs. Lastly, for each DMR, the methylation difference between each of tissue pairs (i.e. pairwise comparisons) was computed and only DMRs that have significant methylation difference (p-value <= 0.01) and the methylation difference is greater than or equal to 0.3 in at least one of the pairwise comparisons are

retained. The scripts for running this pipeline are included as additional files (DDMR_Identification_CpG_mult.r, DDMR_Identification_RegionAnalysis_mult.r, parallel_run_Ziller.py). The results from this script can be found among the additional files (Ziller_et_al_DMR_finding/DMR_final_with_level.tsv)

To statistically infer DMRs that may vary between individuals (i.e., those DMRs used in "Genetic Origins of Methylation Variation"), which the above methodology from Ziller et al.[6] does not, we defined a stochastic model of our methylation data sets in which the observed number of reads supporting methylated and unmethylated cytosines at each position in each sample is drawn from a binomial distribution. In each sample at each cytosine in the CG context there is a single parameter, x_n^i, corresponding to the true fraction of methylated alleles in the population, or the methylation level, where *i* denotes the position of cytosine and *n* denotes the sample. Our null hypothesis is that the methylation level at this position is equal in all samples ($x_n^i = x^i$ for all *n*). Our procedure is designed to test whether the observed data are consistent with the null hypothesis, or alternatively if there is a significant deviation from equal methylation levels.

To do this, first we compute a goodness-of-fit statistic, *s*, which was introduced and validated by Perkins et al[7]. Specifically, we arrange the observed data in an Nx2 table, with one row for each of N samples and a column for reads supporting methylated and unmethylated cytosines respectively. The number of observed reads in sample *n* at position *i* is o_{nj}^i, where *j* = 1 for methylated reads and *j* = 2 for unmethylated reads. The expected number of reads in sample *n* with methylation state *j* under the null hypothesis is e_{nj}^i:

$$e_{nj}^i = \left(\sum_{m=1}^{N} o_{mj}^i\right)\left(\sum_{k=1}^{2} o_{nk}^i\right)/M^i$$

where $M^i = \sum_{n=1}^{N}\sum_{k=1}^{2} o_{nk}^i$ is the total number of reads in all samples. Note that in the case of sites where we initially determined the methylation level was consistent with the background

non-conversion rate (see the Methylation Calling section above), we treat all reads as unmethylated even if they contain a cytosine. The statistic for the goodness of fit is

$$s^i = \sqrt{\frac{1}{2N}\sum_{n=1}^{N}\sum_{j=1}^{2}\left(o_{nj}^i - e_{nj}^i\right)^2}$$

Next, we simulated read count data under our stochastic model assuming the null hypothesis in the following way:

- Set all cell counts in the table to zero
- Randomly select a cell in the table with probability equal to the expected counts divided by the total number of counts in the table ($\frac{e_{nj}^i}{M^i}$). Increment the value in this cell by one.
- Repeat this procedure M^i times.
- Finally, calculate the value of the statistic, s_{shuff}^i, for the randomly generated table.

This randomization procedure was repeated until we observed 100 iterations with a value of s_{shuff}^ithat was at least as extreme as that of the observed data, *s*, up to a maximum of 3,000 iterations. The p-value at position i was then computed as:

$$p^i = \frac{R^i + 1}{T^i}$$

Where R^i is the number of randomized tables with a statistic greater than or equal to the original table's statistic and T^i is the total number of randomized tables that were computed. Our adaptive permutation procedure ensures that any sites which we may potentially identify as significantly differentially methylated with $p^i < 0.01$ will be sampled 3,000 times. At other sites, we have observed an appreciable number (100) of permutations more extreme than our original test statistic ($s \geq s_{shuff}$) and the p-value for these sites will be p ≥ (100+1)/3000 = 0.034; these sites will therefore not be called as differentially methylated.

To control the false discovery rate (FDR) at our desired rate of 1%, we used a computationally efficient procedure designed for comparing multiple sequential permutation-

derived p-values[8]. This procedure is designed to account for the effect of our adaptive permutation procedure on the form of the distribution of p-values. First we generated a histogram of the p-values across all cytosines in CG context. We also calculated the expected number of p-values to fall in a particular bin under the null hypothesis. This expected count is computed by multiplying the width of the bin by the current estimate for the number of true null hypotheses (m_0), which is initialized to the number of tests performed. We then identified the first bin (starting from the most significant bin) where the expected number of p-values is greater than or equal to the observed value. The differences between the expected and observed counts in all the bins up to this point are summed, and a new estimate of m_0 is generated by subtracting this sum from the current total number of tests. This procedure was iterated until convergence, which we defined as a change in the m_0 estimate less than or equal to 0.01. With this m_0 estimate, we were able to estimate the FDR corresponding to a given p-value cutoff by multiplying the p-value by the m_0 estimate (the expected number of positives at that cutoff under the null hypothesis) and dividing that product by the total number of significant tests we detected at that p-value cutoff. We chose the largest p-value cutoff that still satisfied our FDR requirement.

In the next stage of analysis, we combined significant sites (DMSs) into blocks if they were within 250 bases of one another and had methylation changes in the same direction (e.g., sample A was hypermethylated and sample B was hypomethylated at both sites). A sample was considered hypo or hyper methylated if the deviation of observed counts from the expected counts was in the top or bottom 1% of deviations. These residuals were calculated for a position *i* using the following formula for a given cell in row *n* and column *j* of the table:

$$\frac{o_{nj}^{i} - e_{nj}^{i}}{\sqrt{e_{nj}^{i} * (1 - \sum_{m=1}^{N} \frac{e_{mj}^{i}}{M^{i}}) * (1 - \sum_{k=1}^{2} \frac{e_{nk}^{i}}{M^{i}})}}$$

The distinction between hypermethylation and hypomethylation was made based on the sign of the residuals. For example, if the residual for the methylated read count of sample A was positive, it was counted as hypermethylation. Furthermore, blocks that contained fewer than 10 differentially methylated sites were discarded. The DMRs called with this methodology, along with their methylation levels, are in the additional files (https://bitbucket.org/schultzmattd/methylpy and DMR_by_methylpy/DMR_methylpy_matrix).

Benchmark methylpy and other DMR identification methods

To further evaluate the performance of the DMR finder (methylpy) used to find inter-individual DMRs in the section "Genetic Origins of Methylation Variation", methylpy was compared with three published DMR finding methods: BSmooth[9], DSS[10] and MOABS[11]. The test was done on methylome data of adrenal gland samples from individual 2 and individual 3 (AD-2 and AD-3) and two aorta samples from the same individuals (AO-2 and AO-3). Data and code for this benchmark can be download from this link (https://drive.google.com/folderview?id=0B1BhFMhr3HTATjdWLUx3d1ZtZHM&usp=sharing). For BSmooth and MOABS, the default settings were used. For DSS, we used 1% FDR cutoff for calling differentially methlyated locus (DMLs). Then DMLs within 300bp were merged and regions containing at least 3 DMLs were called as DMRs. Note that these two parameters are the same as the default settings in MOABS. Only data of chromosome 1 was used in this analysis.

Methylation Levels

Throughout the paper we refer to the methylation levels of regions in various contexts. Unless otherwise noted, these methylation levels are more specifically weighted methylation levels as defined here[12]. Sites predicted to be unmethylated (based on the binomial test) had their methylation level set to zero.

RNA-seq Analysis

RNA-seq mapping was done using Tophat2[13] with default parameters ("-r 200", "--library-type fr-firststrand") against the human reference genome version hg19. The genomic features were obtained from GENCODE version 14[14]. We used htseq-count to map reads to GENCODE features and generate read counts using (http://www-huber.embl.de/users/anders/HTSeq) using default parameters except "-s reverse".

RNA-seq Expression Quantification

In order to quantify expression levels of each of the annotated genomic feature, we implemented the "cufflinks" module of the Cufflinks suite version 2.1.1[15]. Cufflinks produces FPKM (Fragments per kilobase of feature per million) for each of the annotated features. We used default parameters, except for the use of –-upper-quartile-norm option and --max-bundle-frags as 50,000,000. This extreme limit was set to avoid skipping of regions with several fragments. The default value of 1,000,000 would result in several tissue-specific or highly expressed genes to be labeled as "HIDATA" without an actual FPKM value being reported. Then, we applied quartile normalization to FPKMs, which is described in http://cufflinks.cbcb.umd.edu/manual.html#library_norm_meth. Specifically, we scaled the 75% quartile FPKM of every sample to be the mean 75% quartile FPKM of all samples (i.e., all 36 tissue samples from this study, IMR90, H1, and placenta samples).

RNA-seq Differential Expression Analysis

In order to obtain genes that are differentially expressed across any of the samples in this study, we used htseq-count to map reads to GENCODE features and generate read counts (http://www-huber.embl.de/users/anders/HTSeq) using default parameters except "-s reverse". These read counts were tested for differential expression using the quasi-likelihood F-test (glmQLFTest)[16] implemented in edgeR[17]. In contrast to pairwise comparisons (like case vs control or wild-type vs treatment) this test does not require specifying which groups would be different. The set of genes enriched or depleted in one group compared to an average of all

other tissues was obtained. An FDR cut-off of 0.05 was used to identify differentially expressed genes.

CG DMR Dendrogram

To create the dendrogram shown in Fig. 1c, we first used the cmdscale command from R to perform multidimensional scaling and compute the first 15 principal components of the CG DMR methylation level matrix. The percent variance explained from this multidimensional scaling is presented in Extended Data Fig. 1c. Next, we used the heatmap.2 function in the R package gplots[18] with the default distance metric, and the Ward hierarchical clustering method on these principal components to generate the dendrogram.

Differentially Expressed Genes Dendrogram

To create the dendrogram shown in Fig. 1d, we first used the cmdscale command from R to perform multidimensional scaling and compute the first 15 principal components of the RPKM values, which were first normalized by the maximum expression value observed at each locus, from all differentially expressed genes. The percent variance explained from this multidimensional scaling is presented in Extended Data Fig. 1d. Next, we used the heatmap.2 function in the R package gplots[18] with the default distance metric, and the Ward hierarchical clustering method on these principal components to generate the dendrogram.

Genomic Feature Definitions

Promoters were defined as -1000bp to +300bp region of the transcription start sites of transcripts defined in GENCODE version 14[14]. Exons and introns were also defined using the GENCODE reference. Putative enhancers were obtained from Leung, Rajagopal, and Jung et al.[19] which were predicted using histone mark profiles. CG islands (CGIs) were downloaded from UCSC genome browser[20]. CGI shores were defined as the 2kb regions extending in both directions from CGIs[21,22].

DMR Tissue Specificity Determination

To find CG DMRs that are strongly and specifically hypomethylated or hypermethylated in a particular tissue, we ranked tissues by the methylation level of a CG DMR (from lowest to highest). Then, starting from the tissue with the lowest methylation level, we computed the difference in methylation level between adjacent tissues. Next, we identified the largest difference, and if it was greater than or equal to 0.1, we divided the tissues into two groups (i.e., hypomethylated tissues and hypermethylated tissues). If the hypomethylated group contained ten or fewer tissues, the DMR was classified as a tissue-specific, hypomethylated CG DMR in those tissues. If the hypermethylated group had ten or fewer tissues, the CG DMR was classified as a tissue-specific, hypermethylated CG DMR in those tissues. We ignored other CG DMRs (including CG DMRs with difference less than 0.1 between adjacent ranked tissues) were because their tissue specificity was too obscure.

DMR GO Enrichment

We used GREAT[23] with default parameters to find functional terms of genes near CG DMRs as these terms indicate the potential regulatory functions of these CG DMRs. Since too many DMRs can saturate the Hypergeometric Test it uses, we considered at most the top 5,000 DMRs sample-specific DMRs ranked (largest to smallest) by the difference (which has to be greater or equal to 0.1) in methylation level between the hypermethylated and hypomethylated groups as input. Furthermore, we require each of these DMRs to have at least 4 DMSs. We focused on the GO Biological Process and Mouse Phenotype categories and representative results from this analysis are shown in Extended Data Fig 1e and f. The complete results are in Supplementary Tables 2 and 3.

Correlating Methylation States of DMRs with Gene Expression

To compute the correlations shown in Fig 2a, we used the nearest gene model to predict the target gene of every DMR (i.e., the gene with the closest transcription start site was predicted as the target gene of a DMR). Then, we computed the Spearman correlation coefficient between the methylation level of that DMR and the expression level of its target gene. Only intergenic

hypomethylated DMRs with differentially expressed protein-coding genes as a target were included in this analysis.

To understand the role of these DMRs and their association with expression, we grouped them into different categories according the genomic elements they did or did not overlap. Genebody DMRs were defined as those that overlapped gene bodies. Enhancer DMRs were defined as those that overlapped enhancers. Promoter, CGI and CGI shore DMRs were defined as those that overlapped promoters, CGIs, or CGI shores. DMRs not in these categories and lying outside any gene body labeled as intergenic. Finally, undefined intragenic DMRs were those that didn't overlap any of these categories. As a control we shuffled the sample labels of the methylation levels and computed the Spearman correlation coefficients as above, which labeled as shuffled.

Annotating undefined intragenic DMRs and promoter DMRs

We used K-means clustering to cluster the histone modification profiles of undefined intragenic DMRs (uiDMRs). We assigned the strand of target gene to each uiDMR to ensure that the TSSs of target genes were always upstream of uiDMRs and eliminate the possibility that strandedness would affect the clustering. Next, we divided each uiDMR into 10 equally sized bins and we divided the 5kb region on either side of each uiDMR into 100bp bins. We split DMRs into equally sized bins for several reasons. Firstly, DMRs varied in length, and we needed a way of comparing the locations of motifs in different DMRs. Secondly, to estimate and show the location preference of motifs, DMRs needed to be binned in order to get an appreciable number of motif instances falling to fall in each position across a DMR. Finally, we wanted to avoid splitting DMRs into bins with different sizes to keep the analysis unbiased as we did not want to introduce confounding factors like differing bin sizes in a single DMR.

We then created a vector of input-normalized ChIP-seq RPKMs of the six histone marks for each bin. The uiDMRs were then clustered into five groups using these vectors. We labeled these groups as weak enhancer (strong H3K4me1, depleted H3K4me3 and strong H3K27ac),

promoter-proximal (near region with strong H3K4me3 and strong H3K27ac and depleted in H3K4me1), transcribed (strong H3K36me3), poised enhancer (strong H3K4me1 and weak H3K27ac) and unmarked (no noticeable active histone marks).

We performed a similar analysis for DMRs that overlapped promoters (i.e., the same fixed window definition previously mentioned). Not all of these regions were active (i.e., marked by H3K4me3 and H3K27ac), so to identify active and inactive promoters we applied K-means clustering to the histone modification profile of promoter DMRs into two categories: strong promoters and unmarked promoters. DMRs in strong promoters showed an H3K4me3 and H3K27ac signal; whereas, DMRs in unmarked promoters displayed at most a very weak H3K27ac and H3K4me3 signal.

Sequence motifs enriched in tissue-specific uiDMRs and tissue-specific enhancers were identified using Homer[24].

DNase I sensitivity analysis

To plot DNase I sensitivity data of fetal tissues in Extended Data Fig. 5, we downloaded DNase I data from GEO (GSE18927 and Supplementary Table 6). To profile the DNase I sensitivity of unmarked uiDMRs, we divided each unmarked uiDMR into 10 equally sized bins and the 2.5kb region on either side of each uiDMR into 50bp bins. The DNase I sensitivity RPKMs were calculated for each bin for each unmarked uiDMR, and the values were aggregated to generate the average profile. The same approach was applied to generate the average profiles of DMRs overlapping intragenic enhancers and unmarked uiDMRs with shuffled locations. Only DMRs greater than 200bp in length (i.e., each bin is greater than 50bp) are included in this analysis.

Measuring the genetic origins of DNA methylation

If DNA sequence is involved in regulating DNA methylation we should observed an enrichment of sequence variants where there is epigenomic variation. To rank DMRs by epigenomic variation, we created a tissue-specific methylation outlier score (MOS). The MOS takes

advantage of some tissues methylomes being sequences in triplicate and identifies DMRs where one individual's methylation state is divergent from the other two. MOS is calculated as,

$$\mathrm{MOS}_{i=}\left|\frac{\Delta_{ij}+\Delta_{ik}}{2}\right|-|\Delta_{jk}|$$

, where *i*, *j* and *k* represent the three individuals and Δ_{ij} represents the difference in methylation state scores between individuals *j* and *k*. MOS_i represents the degree to which individual *i* is an outlier at a particular DMR. We subtract $|\Delta_{jk}|$ to account for background level of DNA methylation variability at the DMR. A separate *MOS* is calculated for each individual at each DMR. Each DMR is assigned its single greatest *MOS* score and the corresponding individual is considered the outlier.

We hypothesized that MOS performs better than standard deviation as it considers the level of similarity between the two concordant replicates. Thus, DMRs where variation might be increased by measurement error are less highly ranked as some measurement errors may be consistent across the samples, and therefore, would increase the variation between the concordant replicates. The motif associated SNPs (maSNP) occurrence in the top 2,500 DMRs ranked by standard deviation was: FT = 1.51; GA = 1.45; PO = 1.61; SB = 1.61; SX = 1.46. These numbers result in an average maSNP occurrence of 1.528. When MOS is used to rank the DMRS the enrichment scores are: FT = 1.58; GA = 1.65; PO = 1.65; SB = 1.60; SX = 1.63. The MOS ranked DMRs result in an average maSNP occurrence of 1.622. Thus, MOS does a better job or ranking DMRs by their enrichment with maSNPs.

Further, to determine that maSNP enrichment of DMRs when ranked by MOS was statistically significant we used a Chi-squared test to compare the association between the number of maSNPs and non-maSNPs in a DMR and its SD or MOS rank. To do this, the maSNP and non-maSNP counts were compared between the top MOS ranked 2500 DMRs and the DMRs ranked between 497500 and 500000 (i.e., we constructed a 2x2 table where rows indicated

whether or not the DMR was in the top 2500 DMRs and columns indicated whether or not that DMR contained an maSNP). The P-values for maSNP enrichment in the top 2,500 MOS ranked DMRs were: FT = 0.0006811861; GA = 2.443996e-16; PO = 4.2191e-16; SB = 0.00202069 and SX = 6.313224e-08. Thus, demonstrating the significance of the maSNP enrichment in the MOS ranked DMRs. The Chi-squared test of significance was repeated using DMRs ranked by strand deviation: FT = 0.01908347; GA = 0.09873; PO = 6.997994e-07; SB = 0.0003348352 and SX = 0.002674707. In all cases, the P-value was more significant for the MOS ranked DMRs.

To evaluate the level of sequence variation at *cis*-regulatory elements we created sets of DNA motifs that are putatively involved in the tissue-specific regulation of DNA methylation levels at the DMRs. For each tissue we created two *de novo* motif sets: (i) hypo and (ii) hyper. The tissue-specific *de novo* motif sets were created using the Epigram pipeline[25] to identify a set of motifs that are discriminative of tissue-specific hypo and hypermethylated regions. Briefly, the Epigram pipeline works as the following: (i) the two sets of sequences (tissue-specific hypo and hypermethylated regions) are balanced so that they have the same distribution of lengths and GC-content; (ii) two *de novo* motif finding methods, HOMER[24] and its own, are used to identify motifs that are enriched in either set; (iii) a LASSO logistic regression[26] is used to select the motifs that are most discriminative of the two regions; (iv) a Random Forest classifier and 5-fold cross-validation are used to assess the collective ability of the motifs to classify the sequences into hypo or hypermethylated; (v) a second round of feature selection is performed to heuristically select a subset of 20 motifs that has the greatest discrimination power. Thus, the Epigram pipeline identities motifs that are predictive of tissue-specific hypo- and hypermethylation and measures their ability to distinguish the two sets.

During the creation of both the *de novo* and known motif sets it is necessary to have sets of tissue-specific hypo and hypermethylated regions. The tissue-specific hypomethylated regions were taken from the DMR GREAT analysis as previously defined. The set of hypo- and hyper-

methylated sequence sets were then balanced so that they were equal in size and had the same distribution of GC-content and region lengths[25]. The number of hypomethylated DMRs for each tissue after sampling ranged from 278 to 15,732 with a mean of 7,307 while the hyper sets ranged from 745 to 12,190 with a mean of 6,028.

To create known set of known motifs five motif databases were combined: (i) Transfac[27], (ii) Jaspar[28], (iii) Uniprobe[29], (iv) hPDI[30] and (v) Taipale[31]. We removed known if their name was not listed in GENCODE or they were not annotated with the gene ontology term 'sequence-specific DNA binding' or 'DNA binding'. To make the final set of motifs non-redundant, if there was more than one motif for the same gene, then only the motif with the greatest information content was retained. To calculate motif-breaking cut-offs for the known motifs we created background distribution and took a cut-off that corresponds to a 0.05 *P*-value. Taking the DMR DNA sequences and shuffling them so that order of nucleotides was randomized created the background distribution sequences. A motif specific background distribution was created by recording the best score of *S* (see above) in each of the shuffled sequence.

PMD Identification

To identify PMDs, we created a random forest classifier. Random forests are an ensemble machine learning technique (described in detail here (Breiman, L. Random Forests. Machine Learning. 2001)) used for classification. We first visually classified regions on chromosome 22 that we felt were strong candidates as PMDs or non-PMDs (Supplementary Table 7). These regions were then used to train a random forest, which was implemented in the python function RandomForestClassifier from the module sklearn.ensemble[32]. Specifically, we then divided these regions into 10kb nonoverlapping bins and computed the percentiles of the methylation levels at the CG sites within each bin. We divided genome into 10kb non-overlapping bins mainly to reduce the effect of smaller DNA methylation variation. PMDs were first discovered by Lister et al. as large (mean length = 153kb, PMID: 19829295) regions with intermediate

methylation level (< 70%, PMID: 19829295). Consequently, we chose a large bin size (10 kb) to reduce the effect of methylation variations in smaller scale (such as DMRs). Furthermore, the features (methylation level distribution of CG sites) used in classifier required enough CG sites inside each bin to accurately estimate this distribution, which necessitated a relatively large bin. We excluded 10kb bins with fewer than 10 CG sites because of the same reason mention above: accurately estimating the methylation level distribution of CG sites inside bin required enough number of sites. Therefore, for bins with very few CG sites (< 10 here), we were unable to classify them (into “PMD” or “non-PMD”).

These percentiles were used as features for the random forest. The following arguments were supplied to the Python function:

n_estimators = 10000, max_features=None, oob_score=True, compute_importances=True

In this procedure, out-of-bag error estimation is used to assess the performance of the classifier. More specifically, when building the classifier, the training data can be bootstrap sampled, which leaves a portion of the data out of the classifier’s construction and can later be used to assess the rate at which the classifier is correctly predicting known labels. To assess the performance of our models, we calculated one minus the out-of-bag error rate reported by RandomForestClassifier, which yielded a correct prediction rate of at least 90% (PA-2 - 90.23%, PA-3 - 92.37%, IMR90 - 97.65%, PLA - 92.33%).

Comparing PMDs Called in IMR90, PA-2, PA-3 and Placenta

We used GAT to estimate the significance of the overlap between PMDs in different samples shown in Extended Data Fig. 6c. The workspace we used was the human reference genome (hg19) excluding ENCODE blacklisted regions. The options provided to GAT were: “--ignore-segment-tracks --num-samples=1000 --bucket-size=10000”.

Histone Modification Profiles Across PMDs

To profile the histone marks in PMDs and the surrounding regions shown in Fig. 2e, f, we divided the 300kb upstream and downstream of each PMD into 10kb bins. The body of PMD was divided evenly into 10 bins. Next, we averaged the input normalized ChIP-seq RPKM for each bin. As a control we shuffled the PMDs and performed the same computation.

Testing Histone Modification Enrichment and Depletion Inside and Outside of PMDs

For each histone mark and separately for each sample (PA-2 and IMR90), we grouped the signal medians displayed in Fig. 2e, f by whether they were inside or outside of the PMD. Next, we performed a Mann-Whitney test on these groups to estimate the significance of the difference in signal medians inside and outside of PMDs.

mCH Motif Calling

To find the predominant nucleotide context of mCH in each sample, we took the top 800,000 methylated, mCH sites (the least number of sites in the three samples displayed in Fig. 3b-d) that did not overlap with a heterozygous SNP and input the surrounding (+/- 5bp) nucleotides from the SNP-corrected reference genomes to the seqLogo package in Bioconductor.

Distribution of Expression Across mCH Quantiles

To examine the correlation between expression and mCH, we binned the expression levels genes into quantiles based on the mCH levels in the tissue where expression was measured. For example, the boxplot in Extended Data Fig. 8b labeled “85” contains expression levels from all the genes that were between the 85th and 90th quantile of mCH level. It is important to note that the absolute methylation level for the 85th and 90th quantile will vary from tissue to tissue. We took this approach to account for the differences in cellular heterogeneity between these tissues.

mCH Pattern Clustering

To identify sets of genes that share similar DNA methylation patterns in an unbiased fashion, we applied a procedure that combines dimensional reduction using principal component analysis, followed by clustering (Lister et al., 2013). We profiled the methylation level (mCAC/CAC and

mCAG/CAG) in gene bodies (TSS-TES) and 5' promoter regions (1 kb upstream of the TSS) within each of 25 samples included in this analysis (collapsed tissue replicates, NRN, GLA, H1 and its derivatives). The methylation level in each sample for each gene was normalized by the average over the gene's distal flanking region (50-100 kb upstream of TSS or downstream of TES). Normalized mC/C values were then log-transformed. These data were combined into a matrix of 104 features for each of 17,138 autosomal genes. Any bins with missing data due to insufficient coverage in one of the samples (0.22% of the total) were replaced with the median value of the entire data set. We performed singular value decomposition on this data matrix to identify the linear combinations of methylation features that account for the largest fraction of the total data variance. We retained the top 7 PCs as a low-dimensional representation of robust genomic methylation features, accounting for 70.3% of the total data variance. Next, we used k-means clustering to estimate gene sets with highly similar withinset methylation patterns. We chose to extract k=20 clusters to capture a diverse range of methylation features, while still allowing visualization and statistical enrichment analysis of functional association for each gene set. We repeated the clustering procedure 5 times using random initialization of the cluster centers, choosing as the final estimate the run with the smallest within-cluster sum of distances from each point to the cluster centroid.

To display the methylation patterns within these gene clusters in Fig. 3f, we profiled the methylation level (mCAS/CAS) in bins of size 1 kb starting 100 kb upstream of the TSS and ending 100 kb downstream of the transcription end site (TES). To compare genes with different lengths, we divided each gene body into 10 non-overlapping bins of equal size extending from the TSS to the TES. Methylation levels were normalized by the flanking region as described above. We then linearly interpolated the gene-body mCAS/CAS data at 100 evenly spaced bins within the gene body in order to give roughly equal weight to the gene-body and flanking methylation data. To visualize the heatmaps of mCAS/CAS patterns for each of 17,138 genes,

we smoothed and downsampled the genes 40-fold to allow representation of genome-scale features.

CAC and CAG Correlation Analysis

In Extended Data Figure 8c, d we examined the relationship between mCAC and mCAG in the following way. The total methylation level (mCAC/CAC or mCAG/CAG) was calculated within all autosomal gene bodies (from TSS to TES). We excluded genes shorter than 2kb. We computed the Spearman (rank) correlation coefficient between these two methylation levels across all genes. These correlations may be diminished by noise due to sampling a finite set of reads for each gene. To determine the magnitude of this effect, we simulated MethylC-Seq basecalls under the assumption of a perfect rank correlation of the true methylation levels. The rank correlation of the simulated reads provides an upper bound on the level of correlation that could have been observed.

Read Position Methylation Level Biases

It has previously been noted that sequencing biases may erroneously be interpreted as mCH[9,33]. To test for this possibility, we constructed m-bias plots as described here[9] and found a very slight bias in the methylation level at the beginning of our reads (Extended Data Figure 8f-h). Consequently, we trimmed the first 10 bases of reads in a sample with (PO-2) and without (EG-2) mCH to see if this bias affected our identification of the CAC mCH motif. This analysis revealed that the original and bias-free motifs are highly concordant with the mCH motif becoming slightly stronger in the bias-free sample (Extended Data Figure 8i-l). Given that this gain was so slight, we did not feel it justified discarding roughly 10% of our data, so we proceeded with the untrimmed results.

X Chromosome Inactivation

Gender-specific methylation patterns were examined in 9 pairs of tissue samples from adult male (STL003) and female (STL002), as well as paired neuronal (NeuN+) and glial (NeuN–)

samples from adult male (55yo) and female (53yo)[34]. For each of the genes assayed here[35], we examined the total mCG/CG within the promoter region, defined to be a 1 kb region ending at the TSS, and the total mCG/CG or mCH/CH within the gene body (TSS to TES). For this analysis, we included 612 X-linked genes that were >1 kb in length and met a coverage criterion (>4000 basecalls at CG and CH positions within the gene body in all 22 samples examined). The heatmap in Fig. 4b shows the ratio of gene body mCH/CH in female vs. male, without any correction for the non-conversion rate. The black outline in Fig. 4b indicates genes that were found to be significantly hyper-mCH in female (likelihood ratio test, Yekutieli-Benjamini FDR ≤ 0.15), with at least 1.2-fold greater mCH/CH in female vs. male, and with mCH/CH>0 in the female sample (Fisher exact test, $p<0.01$). The likelihood ratio test takes into account the sample-specific bisulfite non-conversion rate for mCH sites, as calibrated using sequencing of unmethylated lambda phage DNA.

To assess the relationship between female-specific mCH/CH and escape from X-chromosome inactivation (XCI), we relied on a published survey of expression on the inactive human X-chromosome[35]. That study used rodent/human somatic cell hybrids to assign a XCI score to each gene; 0 corresponds to inactivated genes, 9 to escapees, and intermediate values show varying levels of expression from the inactivated X-chromosome. We used liftOver to match 405 of the surveyed genes to our pool of 612 X-linked genes; this set included 34 escapee genes (XCI=9). The box plot (Extended Data Fig. 9a) shows the difference between female and male methylation level for genes ranked according to the X-inactivation status index[35]. For each box, the central black line is the median and the box edges are the 25th and 75th percentiles.

We used receiver operating characteristic (ROC) analysis to assess how well female-specific mCH hypermethylation allows discrimination of X-escapee genes (Extended Data Fig. 9b). The area under the ROC curve (AUC) is a statistical measure of discriminability, which ranges from 0.5 when little or no discrimination information is present to 1 for perfect discriminability. A

similar analysis was done to assess how informative female-specific promoter CG hypomethylation, female-specific promoter mCH hypermethylation and female-specific gene body mCG, respectively, is for predicting X-escapee genes. Results are shown in Extended Data Fig. 9c-e.

Haplotype Reconstruction using HaploSeq

First, genotypes for all donors were obtained as above. Next, Hi-C reads and paired-end genome sequencing reads were mapped independently using Novoalign (http://www.novocraft.com) to the donor variant-masked hg19 genome as described above. We mapped the Hi-C reads as single ends and paired them later using in-house scripts. We then performed GATK walkers such as Indel realignment and base recalibration to obtain high quality mapping. Finally, we combined our high-quality genome sequencing and Hi-C reads and performed HaploSeq[36] to obtain higher resolution haplotypes than using Hi-C data alone. We then improved the resolution of our seed haplotype generated by HapCUT[37] using local conditional phasing. Briefly, local conditional phasing is performed by Beagle (v4.0)[38] using all known variants in the population (1000 Genomes dataset, phase1 v3). Using the seed haplotypes generated by HapCUT, Beagle infers the haplotype of unphased gap variants using a Hidden Markov Model. In order for a variant to be conditionally phased, we required a 100% match between the phase status present in the seed haplotype and the phase status predicted by Beagle.

Allele-specific Mapping of methylome data

We first generated modified references for each sample (STL001, STL002, STL003, and STL011) to avoid biasing mapping towards reads containing the hg19 reference variant. To this end, we used the SNP calls described above and identified high quality SNPs by recalibrating variants using the default parameters of variant recalibration (GATK) (2) and only genotypes of highest quality (100% confidence calls by GATK) were used for downstream analyses. We

masked any heterozygous SNP with a PASS by replacing them with an “N” and replaced any homozygous SNP with the appropriate variant. Using these references, we remapped our methylome data with Bowtie2[1] as this aligner allows for alignment to sequences containing Ns using the default settings with the following modifications: "-k 2","--np 0".

Assigning methylome reads to alleles

Mapped methylome reads were assigned to alleles based on base calls on reads that overlapped phased heterozygous SNPs. For reads overlapping multiple phased heterozygous SNPs, they were assigned to allele with support from majority of phased heterozygous SNPs and reads were discarded if two alleles were with equal support. To assign reads to a particular allele, we used the scripts assign_read_to_allele_WGBS_se.pl found in the “assign_reads” folder (additional files).

Allele-specific methylation analysis

Methylome reads assigned to each allele, were then processed in the same way as that we used for whole sample, which is described above. Then, by comparing methylomes of two alleles, DMRs (i.e. allele-specific methylation (ASM) events) were called using the same approach as described above. We also separated ASM events that were caused by changing one of the alleles cytosine context (i.e., it occurred in one of the two bases following the methylated cytosines) and those that did not. Furthermore, we required that each allele was covered by at least 10 reads. The sequence context of ASM may differ in two alleles and only ASM events that contain CG site(s) in at least one allele were included in following analysis.

Aligning RNA-seq reads to alleles

List of genes showing allele-specific expression in each tissue sample was obtained from Leung, Rajagopal, and Jung et al.[19] Specifically, For RNA-seq data of all tissue samples, the

paired-end reads were mapped using Novoalign to a variant masked transcriptome genome, which was constructed using Useq software based on Gencode annotation (hg18). The mapped reads were assigned alleles according to the sequence match in each variant between two alleles. Then, for each allele, duplicate reads were considered as PCR duplicates and removed with Picard. To determine whether removing duplicate reads in RNA-seq datasets is appropriate during downstream analysis, we investigate the distribution of duplicate reads in terms of gene expression levels. If the duplicate reads are biased to the highly expressed genes the duplicate reads reflect gene expression levels. If not, the duplicate reads can be considered as PCR duplicate reads. We observed that the samples containing high duplicate reads showed uniformly distributed duplicate reads regardless of gene expression levels (data not shown), indicating that the duplicate reads contain a lot of PCR duplicate reads. To avoid any statistical bias during downstream analysis we decided to remove duplicate reads across whole samples.

Although reads were aligned to variant-masked genome, there are still others biases favoring either of alleles. First, to reduce the effect of the mappability bias, we aligned simulated reads spanning surrounding variants location and then checked if one allele was favored than the other. If more than 5% reads were mapped to one allele than the other, those variant loci were removed as they are likely to subject inherent mapping bias. Second, to reduce the effect of copy number variation and allelically biased copy number variable regions on allelic analysis, we compared the coverage between two alleles based on WGS data. Any variant that had more than three standard deviations above the mean coverage of each haplotype was excluded. Any variant showing biased WGS coverage between two alleles was also excluded (binomial test p-value less than 0.05 after Benjamini correction). Lastly, we remove heterozygous variants that were erroneously called during genotyping. The probability of each called heterozygous variant that was actually homozygous was calculated from the likelihood of observing the coverage on

each allele from whole genome sequencing. Only heterozygous SNPs that had a FDR of less than 0.5% were included in downstream analysis.

To identify allelically expressed genes, we performed binomial test (with probability 50% as null hypothesis) on the numbers of aligned reads of two alleles. Only reads spanning exonic regions were counted and only genes containing at least 10 aligned reads were tested. Allelically expressed genes were defined based on 5% FDR cutoff.

Tissue and Individual Variability of Allele-specific Methylation and Expression

We defined an ASM (and ASE) event as individual variable if there was any disagreement across the tissues from a single individual (e.g., FT-1 had an ASM event and SX-1 did not). Similarly, we called a site tissue variable if there was any disagreement across a single tissue from the three individuals (e.g., SB-2 had an ASM event and SB-3 did not).

Association between Allele-specific Methylation and Expression

If there is a strong association between between allele-specific methylation (ASM) and allele-specific expression (ASE) events, we should expect more allelic expressed genes rather than bi-allelic expressed genes are proximal to ASM events. To test this, we calculated the fraction of ASE genes and bi-allelically expressed genes that have at least one ASM event within a certain distance. Bi-allelically expressed genes were defined as genes that were covered by at least 10 reads and whose p-values given by binomial test for allelic expression were greater than 0.2. Then, since the distance between genuine ASM and ASE events was unknown, we varied the distance cutoff from 10kb to 100kb. The computation was done for all samples from triplicate tissues and the aggregated the results are shown in Extended Data Fig. 10b.

Similarly, if ASE is associated with ASM, we should expect more allelic expressed genes can be linked to matched ASM event(s) than matched ASM event(s) with their locations shuffled. Therefore, we computed the fraction of ASE genes that were linked to matched ASM event(s)

and matched ASM events but with their locations shuffled. Similar to analysis above, distance cutoff was varied from 10kb to 100kb. The aggregated the results of samples from triplicate tissues are shown in Extended Data Fig. 10c.

References

1. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).
2. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–1303 (2010).
3. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** (2011).
4. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).
5. Wang, T. *et al.* STAR: an integrated solution to management and visualization of sequencing data. *Bioinformatics* btt558 (2013).
6. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).
7. Perkins, W., Tygert, M. & Ward, R. Computing the confidence levels for a root-mean-square test of goodness-of-fit. **217,** 9072–9084 (2011).
8. Bancroft, T., Du, C. & Nettleton, D. Estimation of false discovery rate using sequential permutation p-values. *Biometrics* **69,** 1–7 (2013).
9. Hansen, K. D., Langmead, B. & Irizarry, R. A. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* **13,** R83 (2012).

10. Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42,** e69–e69 (2014).
11. Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15,** R38 (2014).
12. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet* **28,** 583–585 (2012).
13. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14,** R36 (2013).
14. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22,** 1760–1774 (2012).
15. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
16. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol* **11,** (2012).
17. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40,** 4288–4297 (2012).
18. Bolker, G. R. W. I. R. source code and/or documentation contributed by: B. *et al. gplots: Various R programming tools for plotting data*. (2010).
19. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (2015).
20. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41,** D64–D69 (2013).

21. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* **41,** 1350–1353 (2009).
22. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* **41,** 178–186 (2009).
23. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28,** 495–501 (2010).
24. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38,** 576–589 (2010).
25. Whitaker, J. W., Chen, Z. & Wang, W. Predicting the human epigenome from DNA motifs. *Nat. Methods* (2014). doi:10.1038/nmeth.3065
26. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33,** 1 (2010).
27. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34,** D108–D110 (2006).
28. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38,** D105–D110 (2010).
29. Robasky, K. & Bulyk, M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **39,** D124–D128 (2011).
30. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. & Qian, J. hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics* **26,** 287–289 (2010).
31. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339 (2013).

32. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).
33. Kao, W.-C. & Song, Y. S. naiveBayesCall: an efficient model-based base-calling algorithm for high-throughput sequencing. *J Comput Biol* **18,** 365–377 (2011).
34. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* (2013).
35. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434,** 400–404 (2005).
36. Selvaraj, S., Dixon, J. R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31,** 1111–1118 (2013).
37. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24,** i153–i159 (2008).
38. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194,** 459–471 (2013).