Published online 3 December 2015

Nucleic Acids Research, 2016, Vol. 44, No. 1 ***75–94***
doi: 10.1093/nar/gkv1332

Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes

Hongbo Liu[1,*,†], Xiaojuan Liu[2,†], Shumei Zhang[1,†], Jie Lv[3], Song Li[1], Shipeng Shang[1], Shanshan Jia[1], Yanjun Wei[1], Fang Wang[1], Jianzhong Su[1], Qiong Wu[3] and Yan Zhang[1,*]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China, [2]Department of Rehabilitation, the First Affiliated Hospital of Harbin Medical University, Harbin 150001, China and [3]School of Life Science and Technology, State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin 150001, China

Received May 09, 2015; Revised October 29, 2015; Accepted November 17, 2015

ABSTRACT

DNA methylation is a key epigenetic mark that is critical for gene regulation in multicellular eukaryotes. Although various human cell types may have the same genome, these cells have different methylomes. The systematic identification and characterization of methylation marks across cell types are crucial to understand the complex regulatory network for cell fate determination. In this study, we proposed an entropy-based framework termed SMART to integrate the whole genome bisulfite sequencing methylomes across 42 human tissues/cells and identified 757 887 genome segments. Nearly 75% of the segments showed uniform methylation across all cell types. From the remaining 25% of the segments, we identified cell type-specific hypo/hypermethylation marks that were specifically hypo/hypermethylated in a minority of cell types using a statistical approach and presented an atlas of the human methylation marks. Further analysis revealed that the cell type-specific hypomethylation marks were enriched through H3K27ac and transcription factor binding sites in cell type-specific manner. In particular, we observed that the cell type-specific hypomethylation marks are associated with the cell type-specific super-enhancers that drive the expression of cell identity genes. This framework provides a complementary, functional annotation of the human genome and helps to elucidate the critical features and functions of cell type-specific hypomethylation.

INTRODUCTION

DNA methylation is a key epigenetic marker that is critical for mammalian development and plays an essential role in diverse biological processes, such as X chromosome inactivation, genomic imprinting and cell type-specific gene regulation (1). The identification of cytosine methylation in the early 1970s (2) led to decades of research on the detection and characterization of DNA methylation in gene regulation. DNA methylation/unmethylation mechanisms are common in all tissues/cells. However, different methylome landscapes have emerged from different cell types, even though they possess the same genome (3).

Numerous studies have mapped DNA methylomes across human cell lines and tissues through a variety of techniques (4), and have characterized several classes of DNA methylation patterns in regulatory regions, including CpG islands (5), CpG island shores (6), tissue-specific differentially methylated regions (7,8), differentially methylated imprinted regions (9), partially methylated domains (10) and large hypomethylated regions (11,12). Previous studies have demonstrated that the tissue-specific differentially methylated regions are associated with tissue-specific gene expression (13). However, the results of most studies on methylation dynamics across human cell types are generated at a limited resolution and with small sample cohorts. In addition, the characterization of the roles of DNA methy-

*To whom correspondence should be addressed. Tel: +86 86669617; Fax: +86 86669617; Email: tyozhang@ems.hrbmu.edu.cn
Correspondence may also be addressed to Hongbo Liu. Tel: +86 86669617; Fax: +86 86669617; Email: hongbo919@gmail.com
†These authors contributed equally to the paper as first authors.

lation in cell type-specific gene regulation has been limited by the ability to accurately and comprehensively map a high resolution atlas of the cell type-specific methylation marks (MethyMarks) across human cell types (14,15). Thus, the genomic distribution of cell type-specific MethyMarks across human cell types and the regulatory context of these modifications remain a subject of great interest. Mining the MethyMarks of stem cells, particularly human embryonic stem cells (hESCs), is valuable for exploring the role of DNA methylation in the maintenance of pluripotency.

Cell type-specific phenotypes are defined by complex regulatory networks that are driven by multiple genetic and epigenetic regulators, including DNA methylation and transcription factors; however, these mechanisms remain unclear. Thus, the modelling of genetic networks requires the parsing of the interplay between DNA methylation and other cell type-specific regulators. DNA methylation might affect the binding affinity of transcription factors to transcription factor binding sites (TFBSs) in a transcription factor-specific and cell type-specific manner (16,17). For example, the binding variability of a well-known transcription factor CTCF across human cell types has been associated with differential DNA methylation (18). Moreover, it has been reported that enhancers harboring specific epigenetic marks play important roles in the regulation of cell type-specific gene expression (19). Most recently, Andersson et al. identified and characterized an atlas of cell type-specific active enhancers across human cell types and tissues (20). Richard A. Young and his colleagues produced a catalog of super-enhancers, which are large clusters of transcriptional enhancers that play key roles in human cell identity (21,22). Interestingly, accumulating evidence has shown that cell type-specific enhancer activity is dependent on the DNA methylation status (23,24). However, as a consequence of the currently limited annotation of cell type-specific methylation marks, the models and biological roles of DNA methylation in the regulation of enhancer activity remain underexplored. Together, these studies have underscored the roles of DNA methylation as a defining feature of cellular identity, and the systematic identification and characterization of cell type-specific MethyMarks in different human tissues and cell types are needed.

Bisulfite treatment coupled with whole-genome sequencing (variably termed, BS-Seq, WGBS or MethylC-Seq) has generated the most comprehensive single-nucleotide resolution DNA methylome maps (25). The DNA methylomes across multiple human tissues and cell lines that have been profiled using these bisulfite-based technologies provide us with an opportunity to completely map and dissect the DNA methylation marks for various human cell types (3,10,26–28). Some useful tools have been developed to analyze these large-scale DNA methylomes. For example, the CpGMPs described by Su *et al.* (29) can be used to identify the hyper/hypomethylation regions in a given methylome. RnBeads (30) was developed to analyze large-scale DNA methylation data and identify the differentially methylated regions between two samples. For the predefined genome regions, such as CpG islands or gene promoters, the QDMR technique developed by Zhang *et al.* (31) can be used to identify the differentially methylated regions across multiple cell types. However, there are no tools for the de novo identification of differentially methylated regions and cell type-specific MethyMarks in a large number of DNA methylomes.

Here, we describe a novel entropy-based framework, termed 'SMART' (Specific Methylation Analysis and Report Tool), which is focused on integrating a large number of DNA methylomes for the de novo identification of cell type-specific MethyMarks. Using SMART, we propose a comprehensive atlas of the MethyMarks across human cell types. The systematic analysis of this atlas revealed distinct features of the different methylation patterns at regulatory elements. We identified the various roles of the uniformly hypomethylated and hypermethylated regions across human cell types. Importantly, we identified a large number of cell type-specific MethyMarks that were associated with genes with cell type-specific functions. In particular, cell type-specific hypomethylation might reflect the assembly of cell type-specific super-enhancers that drive the expression of the genes that define cell identity.

MATERIALS AND METHODS

Data collection

All DNA methylation data and the other datasets used in this study are listed in Supplementary Table S1.

Quantification of the methylation specificity for CpG sites across cell types

For each CpG site i, a one-step Tukey biweight (TB_i) was calculated based on the raw methylation values $CpG_i(rm_1, rm_2, \cdots, rm_c, \cdots, rm_N)$ across N cell types as $\sum_{c=1}^{N}[w_c \times rm_c]/\sum_{c=1}^{N} w_c$, where w_c is a weight that is calculated by the bisquare function in each cell-type, as described in our previous study (31). Using this method, the one-step Tukey biweight provides a robust weighted mean that is relatively insensitive to outliers. The raw methylation values are processed by the one-step Tukey biweight as $CpG_i(m_1, m_2, \cdots, m_c, \cdots, m_N)$ where $m_c = |rm_c - TB_i|$ (Supplementary Figures S1 and S2). Then, the methylation specificity of a CpG across multiple cell types was calculated using the normalized Shannon entropy as

$$1 - \left[-\sum_{c=1}^{N} p_c \log_N(p_c)\right] \times \left[1 - \frac{\max(rm_c) - \min(rm_c)}{\mathrm{MAX} - \mathrm{MIN}}\right]$$

where $p_c = m_c/\sum_{c=1}^{N} m_c$, $\max(rm_c)$ and $\min(rm_c)$ were the max and min raw methylation level of region r in all samples, respectively, and the MAX and MIN were defined as the highest methylation level 1 (or 100%; the methylation level ranges from 0 to 100%) and the lowest methylation level 0, respectively. Methylation specificity ranges from 0 for the uniformly methylated regions in all samples to 1 for the specifically hyper/hypomethylated regions in a single sample with the largest range.

To determine the thresholds for methylation specificity, we modelled different methylation patterns by random sam-

pling of different normal distributions with different means and standard deviations and studied the distribution of the methylation specificity. For a given mean methylation level (ranging from 0.0 to 1.0) and a given standard deviation (ranging from 0.0 to 0.5), 50 values were random sampled as the methylation levels in 50 samples of a CpG site. This process was repeated 10 000 times to produce 10 000 CpG sites whose methylation specificity across 50 samples were quantified by the method described above. As shown in Supplementary Figure S3A and B, the methylation specificity increases with the standard deviations, suggesting that our method is accurate for quantifying methylation specificity. Methylation specificity is less than 0.5 when the standard deviation is $\leq$0.1, which is usually regarded as having little effect on gene expression. Methylation specificity is more than 0.75 when the standard deviation is $\geq$ 0.3, which would change the methylation status from one to another. Thus, 0.5 was selected as the threshold for the maximum specificity value for those CpGs with a low methylation specificity state and 0.75 as the threshold for the minimum specificity value for those CpGs with a high methylation specificity state. Meanwhile, the CpGs with methylation specificity between 0.5 and 0.75 are defined as having an intermediate specificity state.

It is known that spermatozoa exhibit a decreased level of global methylation, which is very different than other cell types. Thus, to evaluate the performance of our method for quantifying methylation specificity across different numbers of samples, the methylome in spermatozoa cells was included in each dataset, and other samples were randomly selected from the other 49 human methylomes. Using this strategy, we produced ten datasets with different numbers of samples, which ranged from 5 to 50, in step 5. In the methylation specificity distribution shown in Supplementary Figure S3C, two peaks at approximately 0.25 and 1.0 were found in all datasets, suggesting the coexistence of uniformly methylated CpGs and cell type-specific methylated CpGs. Moreover, the height of the peak at approximately 1.0 increases with the sample number, indicating that the greater methylation specificity was caused by the methylation diversity in a larger number of samples. In addition, we examined the methylation of known regulatory elements, including CpG islands, Refseq genes, lincRNAs, ubiquitous enhancers, cell type-specific active enhancers, and super-enhancers. The CpGs in each segment and the flanking 2 kb regions were searched. A composite plot of the methylation specificity of these CpGs was mapped by R for each type of regulatory element (Supplementary Figure S4A).

Analysis of the methylation similarity between neighbouring CpGs across cell types

To measure the methylation similarity between neighbouring CpGs, the Euclidean distance-based methylation similarity between neighbouring CpGs was calculated as $\sqrt{\frac{1}{N}\sum_{c=1}^{N}\left(m_{c,i}-m_{c,j}\right)^2}$. Meanwhile, we calculated an entropy-based methylation similarity measure between two neighbouring CpGs using the methylation specificity in which the raw methylation values are replaced by the absolute methylation difference

$$CpG_{ij}\left(\left|m_{1,i}-m_{1,j}\right|,\left|m_{2,i}-m_{2,j}\right|,\cdots,\left|m_{c,i}-m_{c,j}\right|,\cdots,\left|m_{N,i}-m_{N,j}\right|\right)$$

between two neighboring CpGs. The lower entropy-based methylation similarity indicates that the methylation patterns of two neighbouring CpGs are more similar. For a random case, 1 million CpGs were randomly selected from all CpGs. For each randomly selected CpG, its methylation levels across 50 samples were distributed randomly. Thus, we obtained a random dataset including 1 million CpGs and their methylation levels across 50 samples. Then, the algorithms were applied to calculate the Euclidean distance-based methylation similarity and entropy-based methylation similarity for both the real and random datasets. As shown in Supplementary Figure S5A and B, the distribution of the two types of methylation similarity in the real methylation data and the random methylation data intersect at 0.2 and 0.6, respectively, which are used as the thresholds for judging whether two neighboring CpGs share the same methylation pattern. In addition, we evaluated the effect of the distance between two neighbouring CpGs to the two similarity measures by comparing the features of 250 and 500 bp (Supplementary Figure S5C–J). Compared to 250 bp, the distance threshold of the 500 bp was appropriate for identifying the long-range regions, although a small percentage (4.2%) of short segments may be ignored.

Genome segmentation algorithm

Continuous scanning was performed on each chromosome to obtain the segments composed of CpG sites with high methylation similarity across all cell types. The first CpG site is assigned as a primary segment that is then continually extended by merging the next CpG site that shares a similar methylation pattern across all cell types with the last CpG site in the current primary segment. The conditions used to judge whether two CpG sites share similar methylation patterns include (i) the same specificity state, (ii) entropy-based methylation similarity $<$0.6, (iii) Euclidean distance-based methylation similarity $<$0.2 and (iv) a distance between them of $\leq$500 bp. If these conditions are not satisfied, the extension of current primary segment is completed, and a new primary segment is continually extended, as described above. As the CpG sites in the same primary segment share an almost identical methylation pattern in the same cell type, the mean methylation of these CpG sites is calculated as the methylation level of the primary segment.

It has been reported that incomplete bisulfite conversion and sequencing errors may result in random errors in the methylation status (32). These random errors may cause disconnection of the primary segments that are localized in close proximity and share similar methylation patterns. Thus, two primary segments are merged into a segment if the following conditions are satisfied: (i) the same specificity state, (ii) entropy-based methylation similarity $<$0.6, (iii) Euclidean distance-based methylation similarity $<$0.2, (iv) a distance between them of $\leq$500 bp and (v) no more than five intervening CpGs between them. The evaluation of the threshold of the intervening CpGs indicated that five intervening CpGs should be useful for merging the primary segments into larger segments, without any effect on the seg-

ment features, including the CpG density (Supplementary Figure S5K-P). Finally, the segments with a length of <20 bp or <5 CpGs are filtered out. According to the specificity state, the remaining segments are further classified into different groups, including high specificity segments (HighSpe), low specificity segments (LowSpe) and intermediate specificity segments (InterSpe).

Identification of the cell type-specific methylation marks

When calculating the methylation specificity, the one-step Tukey biweight was calculated as a robust weighted mean using the methylation levels in the majority of cell types after discounting the outliers in the minority of cell types by a weight w_c that was calculated by the bisquare function. Here, we treated the High/InterSpe segment with outliers in a minority of cell types as the potential cell type-specific MethyMark. The MethyMark that is specifically hypomethylated in a minority of cell types is designated as a cell type-specific hypomethylation mark (HypoMark), and the MethyMark that specifically hypermethylated in a minority of specific cell types is designated as a cell type-specific hypermethylation mark (HyperMark). To identify the HypoMarks and HyperMarks for a given cell type, a statistical method based on a one sample t test was developed. For each High/InterSpe segment, we obtained the methylation levels in the majority of cell types with $w_c \geq 0.5$ and set them as the baseline methylation levels. Then, we performed the one sample t test between the baseline methylation levels and each methylation level in the minority of cell types to examine the significance of the potential cell type-specific MethyMarks. If a High/InterSpe segment shows significantly lower methylation (P value $<1.0\times10^{-10}$ and absolute difference to mean baseline methylation level $\geq$0.3) in a cell type compared to the baseline methylation levels, this segment is termed as a HypoMark for this cell type. Similarly, the High/LowSpe segment with significantly higher methylation in a cell type compared to the baseline methylation levels is termed as a HyperMark for this cell type. Due to the similar methylation patterns in two replicate samples for the same cell type, the intersection of HypoMarks/HyperMarks between two replicate samples were selected as the HypoMarks/HyperMarks for the cell type.

Specific methylation analysis and report tool SMART

To facilitate the specific methylation analysis, the algorithms described above were written in Python and integrated into a Specific Methylation Analysis and Report Tool (SMART) that dynamically integrates multiple methylomes and identifies the cell type-specific methylation marks.

The localization of the methylation segments to different features of the genome

To localize the segments to the Refseq genes downloaded from UCSC (33), each segment was classified into seven categories, including the 2 kb upstream of the transcription start site, 5′ UTR, Coding Exon, Intron, 3′ UTR, 2 kb downstream of the transcription stop site and Intergenic, as described in our previous study (34). To localize the segments to the CpG islands, we calculated the overlap ratio of the segments with the CpG islands using Bedtools (35). If more than 50% of a segment overlapped with the CpG islands, this segment is treated as a CpG island segment. In contrast, if >50% of a segment overlapped with a CpG island shore, which is a region that is 2 kb upstream or downstream from the CpG islands, this segment is treated as a CpG island shore segment. The remaining segments are treated as CpG island desert segments. To localize the segments to the repetitive elements downloaded from UCSC, we calculated the overlap ratio of each segment with the repetitive elements using Bedtools. The box plot and kernel density plot of the overlap ratios for the different classes of methylation segments were mapped by the R package 'vioplot'.

Overlaps of the MethyMarks between cell types

The odds ratio Chi-squared test was used to measure the significance of MethyMark overlap between two cell types. For each pair of cell types, we quantified the number of MethyMarks that were common to both cell types, the number of MethyMarks that were only present in the first cell type, the number of MethyMarks that were only present in the second cell type, and the number of MethyMarks that were present in other cell-types, but not in these two cell-types. These four numbers were used to calculate the odds ratio whose significance is estimated by the Chi-squared test or Fisher's exact test, when the conditions for the Chi-squared test were not met.

Correlation between DNA methylation and H3K27ac in the cell type-specific MethyMarks

The H3K27ac chromatin immunoprecipitation sequencing (ChIP-Seq) datasets from 21 cell types were downloaded from the NIH Epigenomics Roadmap Consortium (36). For each MethyMark, the H3K27ac reads whose centres were localized in the MethyMark were counted for each cell type using Bedtools. The H3K27ac reads per kilobase per million mapped reads was used to represent the density of H3K27ac in a MethyMark. For each cell type, Pearson's correlation coefficient was calculated for the DNA methylation pattern and H3K27ac in the HypoMarks and HyperMarks using R.

The chromatin modifications and gene expression related to the hESC H1-specific MethyMarks

The ChIP-Seq chromatin modification data (including H3K4me1/2/3, H3K27ac, H3K27me3 and input) and the RNA-Seq gene expression data in the hESC H1 cell line were downloaded from ENCODE project (37). Ngs.plot (38) was used to visualize the average profiles and heat maps of the $\log_2$ enrichment ratios of several histone marks and transcription factors versus the DNA input at the HypoMark/HyperMarks based on the ChIP-Seq data, with a fragment length equal to 300 bp; the defaults were used for the other parameters. To examine whether the histone marks are enriched specifically in

HypoMark/HyperMarks, we obtained 3 kb (98.8% of the identified segments were shorter than this) of flanking regions of the HypoMark/HyperMarks and visualized the average profiles and heat maps in these regions. The genes related to the hESC H1-specific HypoMarks/HyperMarks were obtained and the read count per million mapped reads across their bodies and ±3 kb flanking regions were visualized using ngs.plot, with the default parameters.

Overlaps of the chromatin states and hESC H1-specific MethyMarks

The 15 types of chromatin states identified by ENCODE project were downloaded from the UCSC table browser. The chromatin states for the same regulatory elements were merged, and 11 types of chromatin states remained. We obtained the chromatin states whose centres were localized in the H1 HypoMarks or HyperMarks. Radar plots of the relative percentage of each chromatin state were mapped to compare their localization in HypoMarks and HyperMarks.

Enrichment analysis of TFBSs in the cell type-specific MethyMarks

We obtained a set of TFBSs of 161 transcription factors in the human genome, which were derived from a large collection of ChIP-Seq experiments performed by the ENCODE project. A TFBS and a particular HypoMark were considered to overlap if the centre of the TFBS was localized in the HypoMark. The enrichment of the TFBSs for a transcription factor over the HypoMarks of a cell type was calculated using the odds ratio Chi-squared test. Using the NANOG and H1 HypoMarks as an example, we assumed that the co-occurrence events (706, 794) between binding sites of all Transcription factors and HypoMarks in all cell types were the background. Then, we counted the number of co-occurrence events (125) between NANOG TFBSs and H1 HypoMarks, with that (55) between the NANOG TFBSs and HypoMarks of other cell types, that (4201) between other TFBSs and H1 HypoMarks and that (702, 413) between other TFBSs and HypoMarks of other cell types. The Chi-square test was performed on a 4-fold table of these four numbers to evaluate the odds ratio (380.00) and its significance ($P < 10^{-100}$). The enrichment analysis of the TFBSs in the cell type-specific HyperMarks was performed in the same way. A heat map of the odds ratio of the uniform TFBSs in the cell type-specific HypoMarks and HyperMarks was visualized by GenePattern (39). For each cell-type, only the top four TFBSs based on the enrichment odds ratio were selected for the heat map. The order of the TFBSs in rows of the matrix was determined by hierarchical clustering, with the distance measured as city-block distance.

Enrichment of the known transcription factor motifs in the MethyMarks

For each cell type, the location of the transcription factor binding sites and motif enrichments in the HypoMarks/HyperMarks were determined using the Homer tool (40) and the default parameters. The known motifs used in this study were derived from the Homer tool when the enrichment P value < 0.05.

Construction of the transcription factor and MethyMark collaboration network in hESCs

The binding sites of 50 transcription factors in the hESC H1 cell line were downloaded from UCSC and mapped to the hESC H1-specific MethyMarks when their centres were localized in a MethyMark. Based on these binding events, a transcription factor and MethyMark collaboration network was constructed and visualized by Cytoscape (41).

Overlaps of the MethyMarks with previously described super-enhancers

The super-enhancers used in this study were obtained from a previous study by Hnisz et al., who identified super-enhancers from 86 human cell and tissue samples based on the H3K27ac ChIP-Seq data (22). For each of the 21 common cell types in this study, we identified the MethyMarks, 50% of which were overlapped with a super-enhancer at least in one cell type. For each cell type, we identified the HypoMarks that only overlapped with the super-enhancers from the same cell type as SuperHypoMarks. The genes related to the hESC H1-specific SuperHypoMarks were also obtained. For each SuperHypoMark, the nearest Refseq or GENCODE gene (version 19) or lincRNAs transcript identified by Cabili *et al.* (42) with distance <2 kb from the SuperHypoMark were identified as SuperHypoMark genes. Moreover the SuperHypoMark transcription factor genes were determined based on the list of transcription factor genes obtained from the study by Vaquerizas *et al.* (43).

Hierarchical clustering and heat maps

Methylation *K*-means hierarchical clustering of the HighSpe segments was performed by Cluster 3.0 (44). *K* was set as different values, such as 6, 8 and 10, to avoid the bias induced by the initial parameter, and the distance measure was Pearson's correlation. The clustering result was viewed in TreeView 1.60 using the default parameters. The heat map view of the DNA methylation and H3K27ac in the cell type-specific HypoMarks and SuperHypoMarks were viewed in TreeView 1.60 using the default parameters. For each cell type, the order of the cell type-specific HypoMarks or SuperHypoMarks in the rows was determined by the methylation level, from low to high. The H3K27ac reads per kilobase per million mapped reads was used to represent the density of H3K27ac in a SuperHypoMark. The box plot and kernel density plot of the DNA methylation and H3K27ac were mapped by the R package 'vioplot'.

Function enrichment analysis for gene sets and genomic regions

All function enrichment analysis for the gene sets (HighSpe segment genes and HypoMark genes) were performed in DAVID using the default parameters (45). The selected genes in each cell type were imported into DAVID to

perform a function enrichment analysis of these genes in the up-expressed tissues, biological processes and KEGG pathways. Moreover, GREAT (46) was used to perform the function enrichment of genome regions, such as the hESC H1-specific HypoMark and cell-type-specific Super-HypoMarks in H1, Hippocampus middle, Gastric and Thymus. The genomic regions were assigned to nearby protein-coding genes based on the basal plus extension rule for regulatory regions (proximal: 5 kb upstream, 1 kb downstream, plus distal up to 500 kb). The annotated terms selected from the enrichment analysis were significant by both hypergeometric and binomial tests ($P < 0.05$).

Determination of the sequence conservation levels

The SiPhy (47) algorithm and software package were used to estimate ω, the deviation of the branch length compared to the neutral tree based on the total number of substitutions estimated from the alignment of the region of interest across 29 placental mammals (build hg19, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/maf/). For global conservation, we estimated ω for each methylation segment produced by SMART based on the human methylomes. The cumulative frequency of conservation levels (ω metric) were mapped for all of the different methylation elements identified in this study.

Data availability

SMART is an open source software. The parameter sets, instructions, and sample data sets, are available at http://fame.edbc.org/smart/. SMART has been released as a Python package called 'SMART-BS-Seq' and is freely available from the Python Package Index (https://pypi.python.org/pypi/SMART-BS-Seq). All resources produced in this study are publicly available through the Human Methylation Mark Atlas (http://fame.edbc.org/methymark).

RESULTS

Methylation specificity analysis tool based on the distance-dependent methylation similarity between neighbouring CpGs

Initially, we integrated 50 existing DNA methylomes in 44 human tissues/cells from the NIH Epigenomics Roadmap Consortium using WGBS and obtained ∼17 million CpGs that were shared by these methylomes (Supplementary Table S2). In each cell type, these CpGs showed bimodal methylation patterns, most of which were hypermethylated (Supplementary Figure S1). For each CpG, the normalized Shannon entropy was used to quantify the methylation specificity across multiple methylomes (Supplementary Figures S2 and S3). The bimodal distribution of the methylation specificity suggested the coexistence of uniformly methylated CpGs and cell type-specific methylated CpGs (Figure 1A). The methylation specificity of known regulatory elements confirmed the high accuracy of this method, such as the low methylation specificity in CpG islands and high methylation specificity in gene promoters (Supplementary Figure S4). Additional analyses of the methylation similarity between neighbouring CpGs based on the Euclidean distance and entropy indicated that the neighbouring CpGs shared similar methylation patterns in all cell types (Supplementary Figure S5A and B). However, this similarity was less obvious when the distance was longer than 500 bp (Figure 1B and Supplementary Figure S5C–J). The distance-dependent methylation similarity between neighbouring CpGs benefits the identification of genomic regions comprising uniformly or cell-specific methylated CpGs.

Based on this feature, we developed a procedure for identifying and characterizing sets of genome segments comprising continuous CpGs with similar methylation specificities (Materials and Methods). For a given set of multiple methylomes profiled using BS-Seq, the entropy-based procedures facilitate the quantification of the methylation specificity for each CpG and the determination of the Euclidean distance and similar entropy for each pair of neighbouring CpGs. Subsequently, continuous scanning, based on these quantified parameters, segments the genome into primary segments comprised of CpG sites with high methylation similarities across all cell types (Figure 1C). Furthermore, the primary segments in close proximity that share similar methylation patterns are merged into larger segments of different types, including HighSpe, InterSpe, and LowSpe segments. Eventually, a statistical method-based one sample t test is used to identify the cell-type-specific MethyMarks from High/InterSpe segments. To facilitate the mining of the MethyMarks across cell types and species, all of the algorithms used in this procedure were integrated into a Specific Methylation Analysis and Report Tool (SMART), which is available at http://fame.edbc.org/smart.

Human genome segmentation based on the DNA methylomes across multiple cell types

The segmentation of the human genome using SMART based on 50 human DNA methylomes identified 757,887 methylation segments covering ∼8.5 million CpGs and ∼538 million bp (Table 1, and Supplementary Table S3). Among these segments, 5406 segments spanned large (≥3.5 kb) chromosomal regions, and 288 segments contained >150 CpGs. As an extreme, the longest segment (chr10:39103301–39130657) included 27 kb and 449 CpGs in a partially methylated domain from the IMR90 cell line and primary spermatozoa from the testis (Supplementary Figure S6). The segments were classified into three groups (HighSpe, InterSpe and LowSpe) that exhibited different features in terms of their length, CpG number, mean/median methylation, methylation specificity, and CpG density (Supplementary Figure S7A-F). In addition, we found that the segments with a low CpG density displayed a high methylation specificity and mean methylation level (Supplementary Figure S7G-I).

Most (∼75%) of the identified methylation segments were LowSpe segments that were uniformly methylated across the 50 methylomes, suggesting the relative stability of DNA methylation in the human genome across multiple cell types (Figure 1D). The distribution of the methylation levels of these segments showed five peaks. The two peaks at ∼0.75 are close to each other and smaller than other three peaks

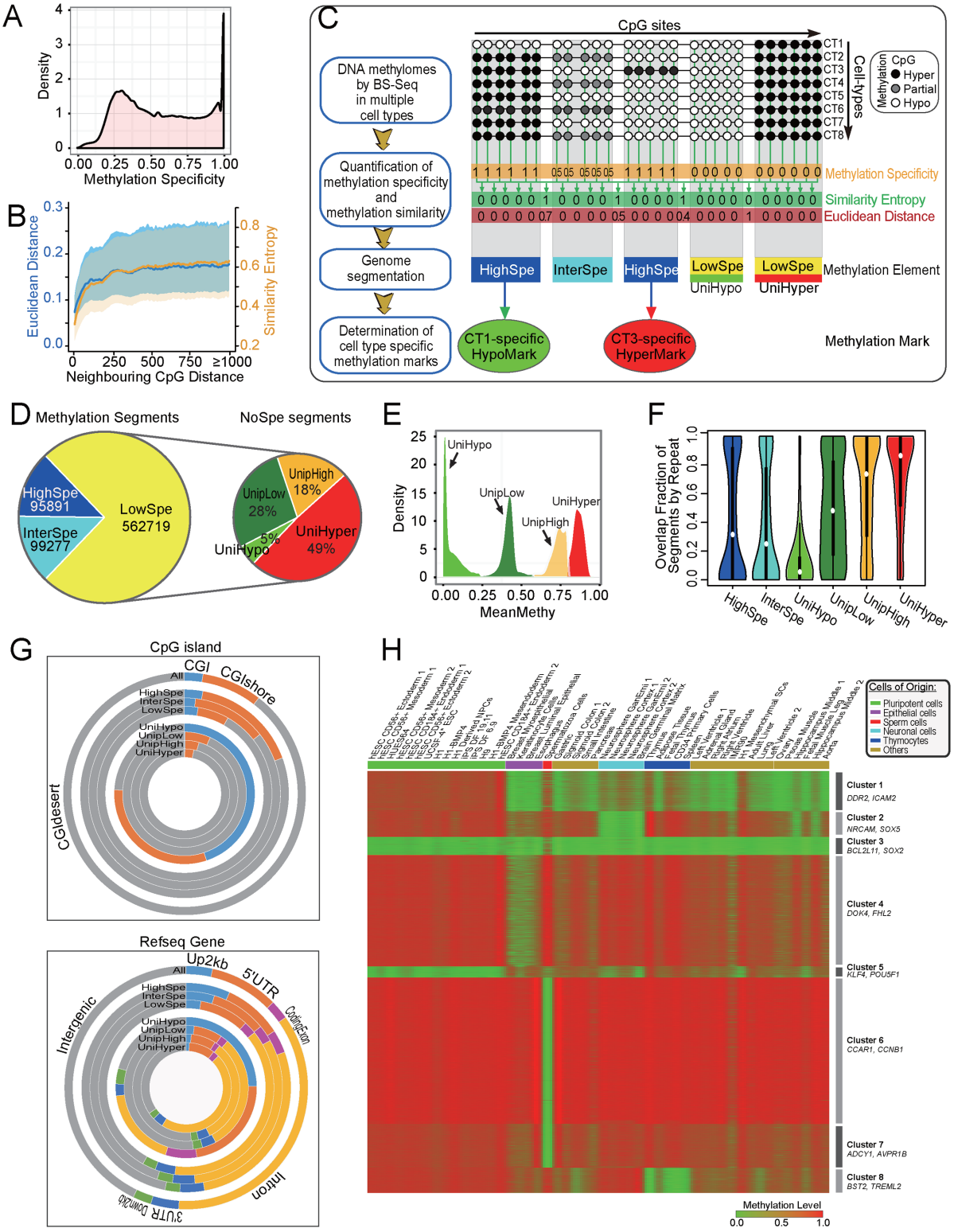


Figure 1. Methylation specificity analysis tool and Human genome segmentation. (**A**) Distribution of methylation specificity across 50 methylomes of CpGs. Methylation specificity of a CpG across multiple cell types ranging from 0, which indicated the uniform methylation in all samples, to 1, which indicates specifical hyper- or hypomethylation in a single sample with the widest range. (**B**) Composite plot of the Euclidean distance and similarity entropy of the DNA methylation levels for neighboring CpGs with different distances. The blue and orange lines indicate the median of the methylation specificity of the Euclidean distance and similarity entropy, respectively. The areas indicate the 25th and 75th percentiles. (**C**) Overview of the methylation specificity analysis framework for genome segmentation based on the methylomes in multiple cell types and determination of cell-type specific methylation marks, including the quantification of the methylation specificity, similarity entropy and Euclidean distance. (**D**) Pie chart for the number of different types of segments in human genome. (**E**) Distribution of mean methylation levels for the LowSpe segments of each cluster across all cell types, as described Supplementary

Table 1. Abbreviation and number of various methylation segments identified in this study

Abbreviation	Description	Number
All	All human segments identified by SMART	757, 887
HighSpe	Segment with high methylation specificity	95, 891
InterSpe	Segment with intermediate methylation specificity	99, 277
LowSpe	Segment with low methylation specificity	562, 719
UniHypo	Uniformly hypomethylated segment in all examined cell types	29, 379
UnipLow	Uniformly partial-low-methylated segment in all examined cell types	154, 919
UnipHigh	Uniformly partial-high-methylated segment in all examined cell types	103, 880
UniHyper	Uniformly hypermethylated segment in all examined cell types	274, 541
MethyMark	Cell type-specific methylation mark that is specifically hypo/hypermethylated in a minority of cell types	460, 438
HypoMark	MethyMark specifically hypomethylated in a minority of cell types	403, 385
HyperMark	MethyMark specifically hypermethylated in a minority of cell types	57 053
Large MethyMark	MethyMark spanning large chromosomal regions of longer than 3.5 kb in length	444
SuperHypoMark	HypoMark only overlapped with super-enhancers from the same cell type	4222

(Supplementary Figure S8A). As the methylation difference between these two peaks is ∼0.05, which is usually regarded as meaningless in methylation analysis, we treated the two peaks as having the same methylation state, partial-high-methylation. Although these two peaks for partial-high-methylation are also close to the peak at ∼0.9, the segments related to these two peaks were not clustered into the same cluster with those related to the peak at ∼0.9 when we performed *k*-means ($k = 3$, 4 and 5) clustering (Supplementary Figure S8B–F). Thus, we used 4-means clustering and classified the LowSpe segments into four categories, including uniformly hypomethylated (UniHypo, 0.00–0.25), uniformly partial-low-methylated (UnipLow, 0.25–0.60), uniformly partial-high-methylated (UnipHigh, 0.60–0.80) and uniformly hypermethylated (UniHyper, 0.80–1.00) segments (Figure 1E). These results revealed that the UniHyper segments accounted for nearly half of the total LowSpe segments, and these segments were located in repetitive elements (Figure 1F). In contrast, the UniHypo segments accounted for only 5% of the total LowSpe segments and were likely localized in CpG islands and gene promoter regions (Figure 1G and Supplementary Figure S9A–D). In addition, these UniHypo segments showed higher levels of the active chromatin marker H3K4me3 than the other types of LowSpe segments (Supplementary Figure S9E). Additional analyses revealed that the ubiquitous enhancers showed low methylation specificity and overlapped with the UniHypo segments (Supplementary Figure S10A and B). Moreover, 312 genes associated with 223 UniHypo segments overlapped with the ubiquitous enhancers, including 66 well-known housekeeping genes (such as *CTCF*) that were enriched in fundamental biological processes and metabolic pathways (Supplementary Figures S10C-E and S11).

Approximately 13% (95, 891) of segments were identified as HighSpe segments, which showed higher specificity than their flanking sequences (Figure 1D and Supplementary Figure S12A). The tissue-specific differentially methylated regions across human tissues/cells that were identified by previous studies showed the same results, thus confirming the reliability of the methylation segments identified in this study (Supplementary Figure S12A). To investigate the effect of DNA methylation on cell identity, we mapped the samples by principal component 1 and principal component 2 obtained from the principal component analysis of the HighSpe segments in the 50 methylomes, and found a distinct methylation pattern in spermatozoa and the clustering of pluripotent cell lines (Supplementary Figure S12B). In further support of these findings, the *k*-means ($k = 6$, 8 and 10) clustering based on the HighSpe segments in the 50 methylomes also revealed that the cell types from similar developmental stages or organ sources shared similar methylation patterns (Figure 1H and Supplementary Figure S13). For example, most of the HighSpe segments were hypermethylated in the pluripotent cell lines, including multiple hESCs, hESC-derived cells and induced pluripotent stem cells (iPSCs), which was distinct from those in other cell types. Using the result of 8-means clustering for example, the HighSpe segments in each cluster showed distinct hypomethylation patterns in specific groups, such as Cluster2 for neuronal cells, Cluster4 for epithelial cells, Cluster5 for pluripotent cells, Cluster6 and Cluster7 for spermatozoa cells, and Cluster8 for thymocytes. Additional analyses on the genomic location and function enrichment revealed that the segments in each cluster were prone to be localized to nearby genes with functions associated with the specific cell types (Supplementary Table S4 and Supplementary Figures S14 and S15). For example, the High-

←

Figure S8. (**F**) The box plot and kernel density plot of the overlap ratio of segments by Repetitive elements. The fraction of the LowSpe segments that overlap with the repetitive elements increases with the DNA methylation level, and the UniHyper segments showed the most overlap with the repetitive elements. (**G**) Genomic localization of the different types of segments relative to the CpG islands (CGI), CpG island shores (CGIshore) and seven refseq gene-related categories, including the 2 kb upstream of the transcription start site (Up2kb), 5′ UTR, Coding Exon (CodingExon), Intron, 3′ UTR, 2 kb downstream of the transcription stop site (Down2kb). (**H**) *K*-means clustering for the DNA methylation patterns of the HighSpe segments in 50 cell types. The samples in six main groups, including Pluripotent cells, Epithelial cells, Sperm cells, Neuronal cells, Thymocytes and Others, were differentially colored. Two examples of the genes associated with each cluster are listed. The methylation level was represented by a gradient from green (unmethylation) to red (full methylation). A larger version of this figure is available in Supplementary Figure S13.

Spe segments of Cluster5 were specifically hypomethylated in pluripotent cell lines, and the associated genes, including the well-known pluripotency factor genes *POU5F1* (also known as *OCT4*) and *KLF4*, were associated with functions for embryonic development and transcriptional regulation. Moreover, those genes related to the HighSpe segments in Cluster2 were specifically hypomethylated in neuronal cells (including *NRCAM* and *SOX5*) and involved in neuron differentiation and neuron development. In agreement with previous findings, the terms for the immune system, including leukocyte/leukocyte activation and immune response were specific for the genes related to the HighSpe segments in Cluster8, which were specifically hypomethylated in immune cell types, including the thymocytes, primary CD34 cells and splenocytes. The *k*-means ($k = 6$, 8 and 10) clustering for the 99 277 InterSpe segments were similar to those of the HighSpe segments (Supplementary Figure S16). These results strongly support the idea that cell type-specific methylation distinguishes human cell types according to their cell type-specific functions, suggesting that the High/InterSpe segments might be potential methylation markers for human cell types.

An atlas of the cell type-specific methylation marks across human cell types

Using SMART, we further identified the HypoMarks and HyperMarks for each of the 42 human cell types and presented a human methylation mark atlas, which is available at http://fame.edbc.org/methymark. This atlas represents the combination of 460, 438 MethyMarks across all cell types and constitutes ∼92% (179, 911) of the total High/InterSpe segments (Figure 2A). Most of the MethyMarks are specifically hypo/hypermethylated in a minority of cell types, and nearly half (83, 683) of these MethyMarks were hypomethylated in a specific cell type. The number of MethyMarks varies considerably across cell types, ranging from 1,000 MethyMarks in sigmoid colon to 68, 381 MethyMarks in the spermatozoa in the testis (Figure 2B and Supplementary Table S5). Although the total number of HypoMarks was ten times higher than the total number of HyperMarks, the percentage of HypoMarks among the MethyMarks in specific cell types ranged from 12.7% in hESC-derived CD184-positive endoderm to 99.2% in the right atrium. Consistent with k-means clustering shown in Supplementary Figure S17, the cell types from the similar developmental stages or tissue sources share a greater number of common MethyMarks (Figure 2B). Additional genomic localization analyses revealed that the cell type-specific HyperMarks are more likely to be located in CpG island-related regions, including CpG islands, CpG island shores and gene promoters ($P < 1.0 \times 10^{-100}$, Chi-squared test) (Supplementary Figure S18A and B). In addition, the MethyMarks significantly overlapped with cell type-specific enhancers ($P < 10 \times 10^{-100}$, Chi-squared test) (Supplementary Figure S18C). Additional enrichment analyses revealed that the genes with promoter HypoMarks in specific cell types are highly expressed in the corresponding cell type and significantly enriched in the corresponding biological functions (Supplementary Figure S18D). These results suggested the existence of cell type-specific MethyMarks and revealed roles for these markers in regulating cell type-specific gene expression and biological functions.

The cell type-specific MethyMarks span large chromosomal regions

Some of the identified cell type-specific MethyMarks spanned large chromosomal regions of longer than 3.5 kb in length (Supplementary Figure S19A). The longest MethyMark spanned 18.5 kb and showed a sperm-specific hypomethylation pattern (Supplementary Figure S19B). These large cell type-specific MethyMarks included 220 HypoMarks (123 unique related segments with 68 Refseq genes) and 224 HyperMarks (65 unique related segments with 42 Refseq genes). As shown in Figure 2C, each of 39 cell types possessed at least one gene associated with large MethyMarks. For example, one MethyMark spanning 3.6 kb was specifically hypermethylated in pluripotent cells and localized to the *PCDHB11* gene, which showed reduced expression in hESC H1 cells compared to brain tissue (Supplementary Figure S20A). In contrast, another MethyMark spanning 4.6 kb was specifically hypomethylated in all pluripotent cells and localized to *AC005062.2* (also known as *LOC101927668*), which is a highly expressed hESC H1-specific long noncoding RNA (Supplementary Figure S20B). Moreover, we also identified several iPSC-specific MethyMarks associated with several genes, including *IRX2*, *C5orf38*, *IRX1* and *SHOX*. For example, the large MethyMarks associated with two Iroquois homeobox genes, *IRX2* and *IRX1*, showed iD19.11-specific hypermethylation, likely reflecting the low expression of these two genes in the iPSC (Figure 2D, Supplementary Figures S20C and S21). Notably, ∼10% of these large MethyMark genes are imprinted genes, such as *MEG3*, *GNAS*, *MEST*, *RB1*, *ZFAT*, *HOXB3*, *VAX2*, *MEIS1*, *HOXC9*, *OTX1* and *EGFL7*, compared to only ∼1% (317) of human imprinted genes (48). For example, a MethyMark in the promoter of the well-known maternally expressed imprinted gene *MEG3* is specifically hypermethylated (>0.9) in pluripotent cells, hypomethylated (close to 0) in primary spermatozoa and intermediately methylated (∼0.5) in other cells and tissues, including the brain, likely reflecting an allele-specific methylation pattern (Supplementary Figure S20D).

The cell type-specific HypoMarks are enriched through H3K27ac

To further characterize the features of the cell type-specific MethyMarks, we explored the states of the well-known histone modification H3K27ac on the cell type-specific HypoMarks and HyperMarks across 21 cell types using the available methylation and H3K27ac data. Nearly half (46%) of the 75, 651 cell type-specific HypoMarks identified in these cell types displayed specific hypomethylation in only a single cell type (Figure 3A). Interestingly, these cell type-specific HypoMarks showed high levels of H3K27ac in the corresponding cell type, while the HyperMarks showed low levels of H3K27ac (Supplementary Figure S22). The correlation analysis revealed that the DNA methylation of both HypoMarks and HyperMarks is significantly negatively correlated with H3K27ac in all cell types (Supplementary Figure S23). Moreover, we found that DNA methylation and

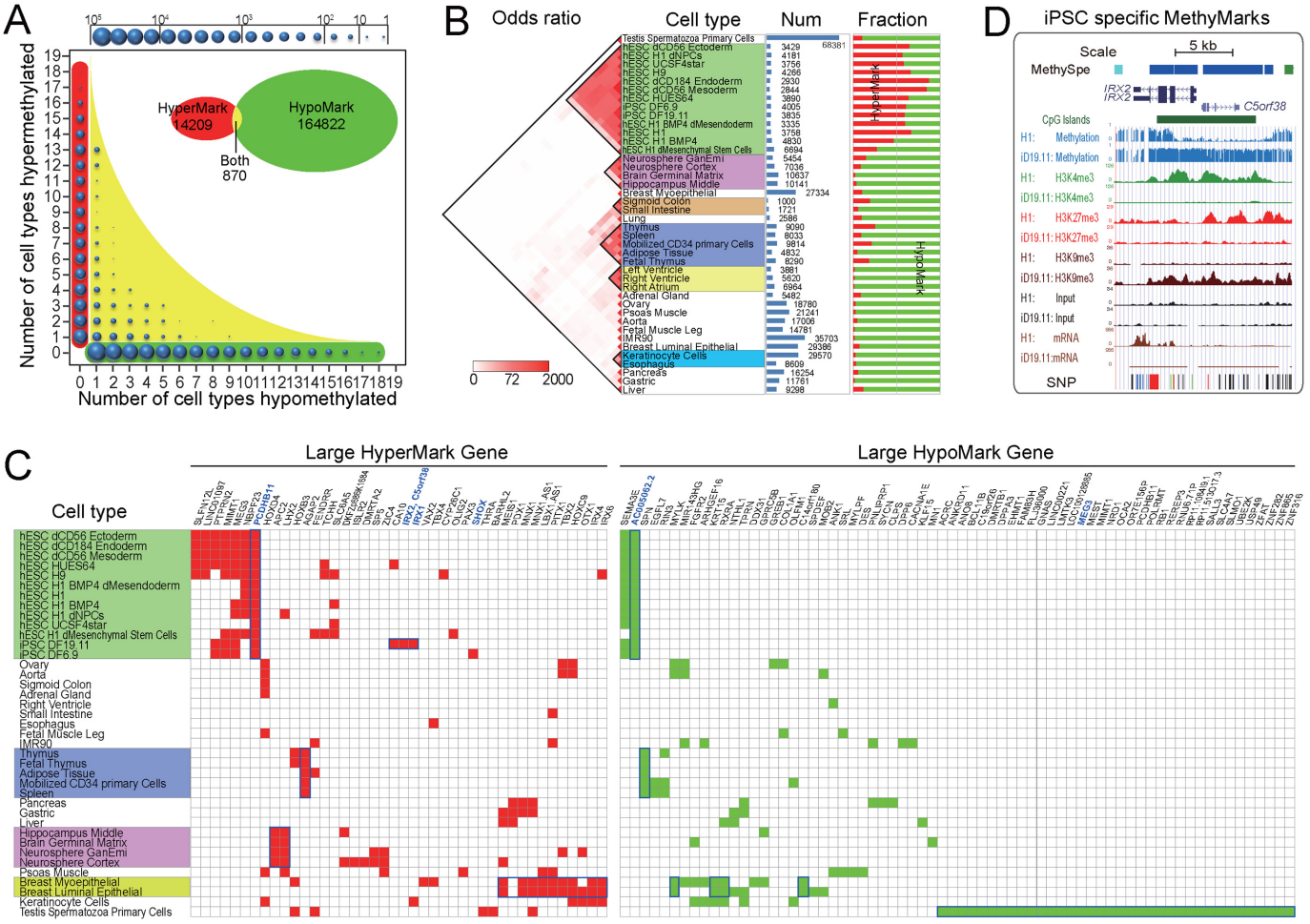


Figure 2. An atlas of the cell type-specific methylation marks across human cell types. (**A**) The cell type specificity of the HypoMarks and HyperMarks. Each cell represents the number of MethyMarks identified as HypoMarks (Column) or HyperMarks (Row) in the corresponding number of cell types. (**B**) The features of the MethyMarks in 42 cell types. The three panels show the overlap of the MethyMarks between cell types, and the number of MethyMarks and the HypoMark/HyperMark fraction in each cell type. (C) Genes related to the large MethyMarks in 39 cell types. The colored grids represent the large HyperMarks (red) or HypoMarks (green) in each cell type. The gene names are listed on the top. (**D**) An example of the iPSC-specific HyperMarks that were associated with the *IRX2* gene.

H3K27ac showed a stronger correlation with the hESC MethyMarks than other genomic segments (Figure 3B and Supplementary Figure S24A). Regarding other chromatin modifications, we observed the enrichment of H3K4me1 (a well-known enhancer marker), H3K4me2 and H3K4me3 (well-known promoter markers) in both HypoMarks and HyperMarks (Figure 3C and Supplementary Figure S25A and B). In addition, we did not find a strong correlation between DNA methylation and H3K27me3 in the hESC MethyMarks (Supplementary Figure S24B). The specific enrichment of H3K27ac in the HypoMarks suggested a reciprocal interaction between DNA hypomethylation and H3K27 acetylation, which may be essential to separate the active enhancers from the poised enhancers marked with H3K4me1 (49). To confirm this hypothesis, we calculated the preferred localization of the hESC H1-specific HypoMarks and HyperMarks in the different chromatin states learned based on a multivariate Hidden Markov Model in hESCs (50). We observed that both the HypoMarks and HyperMarks showed an increased overlap with weak enhancer states, but only HypoMarks overlapped with strong enhancer states, which activate the transcription of nearby genes (Figure 3D). Correspondingly, the genes with promoter HypoMarks showed higher transcript abundance than those with promoter HyperMarks (Figure 3E). These results indicated the distinct roles of the cross-talk between hypomethylation and H3K27ac in the cell type-specific HypoMarks to recruit and form active enhancers that mediate both the spatial and temporal control of development by activating and/or repressing transcription in specific cells.

The cell type-specific HypoMarks frequently co-localize with cell type-specific TFBSs

To determine whether the HypoMarks and HyperMarks facilitate transcription factor binding to DNA sequences in a cell type-specific manner, we obtained the TFBSs (n = 4, 380, 444) of 161 transcription factors from a large collection of ChIP-Seq experiments performed by the ENCODE project. The enrichment odds ratio of each transcription

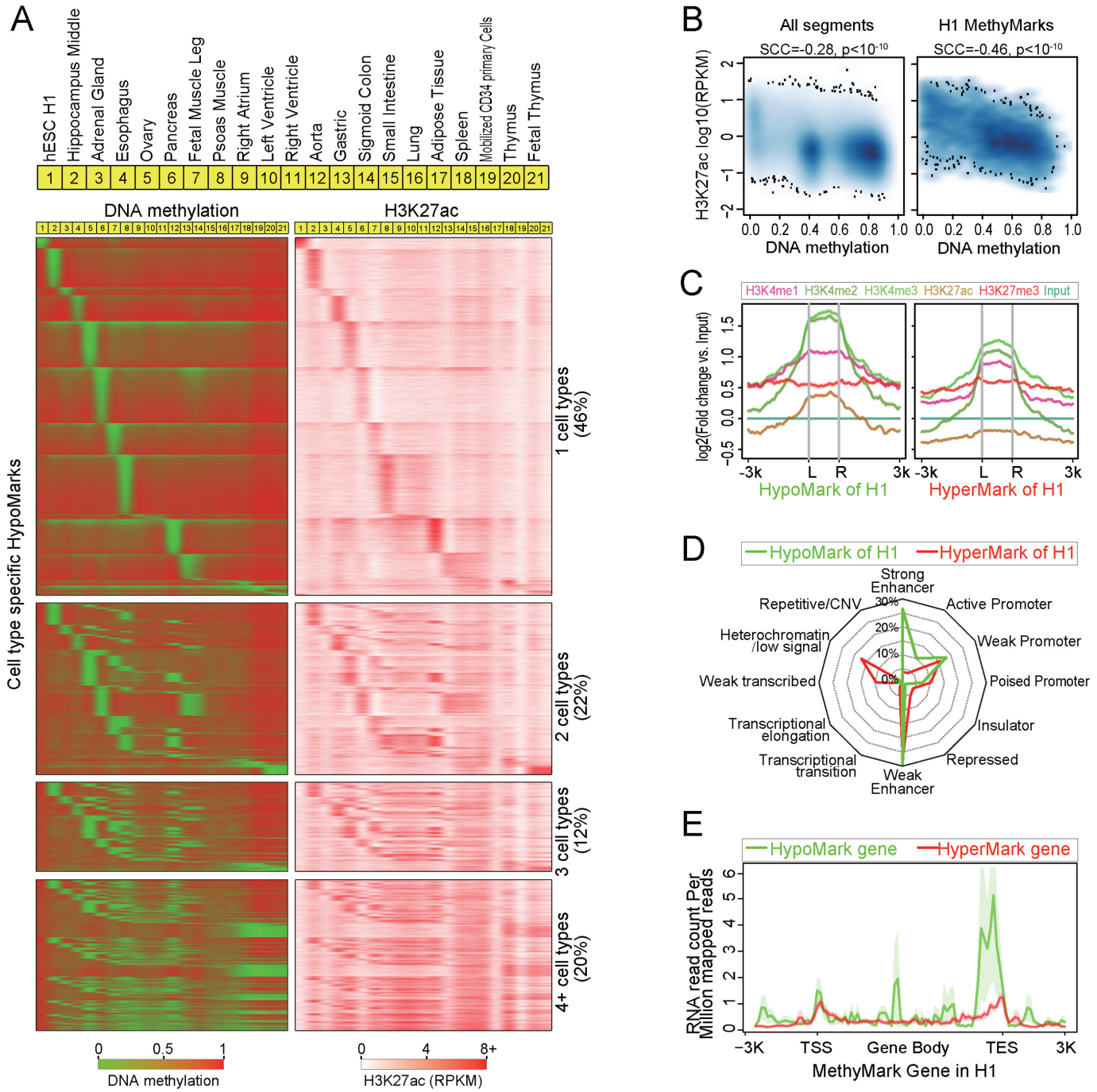


Figure 3. The chromatin modifications and gene expression associated with the cell type-specific HypoMarks. (**A**) Heatmap of DNA methylation and H3K27ac in the cell type-specific HypoMarks. Each row denotes a HypoMark and each column indicates a cell type. The DNA methylation level is represented by a gradient from green (ummethylation) to red (full methylation), and H3K27ac from white (lowest) to red (highest). RPKM represents the H3K27ac reads per kilobase per million mapped reads in a given segment. (**B**) Density scatterplot of DNA methylation and H3K27ac in all segments and H1 MethyMarks. SCC represents the Spearman's rank correlation coefficient calculated by the R function 'cor.test' between DNA methylation and H3K27ac, and p represents the significance of the coefficient. (**C**) Average enrichment profiles of the $\log_2$ ratios of several histone marks and transcription factors versus the DNA input at the HypoMark/HyperMark ±3 kb regions. The lower coordinates 'L'-genomic left and 'R'-genomic right are indicated to the left of higher coordinates. (**D**) Radar plots showing the percentage of the 11 types of chromatin states that overlapped with the HypoMarks (green) and HyperMarks (red). (**E**) The expression levels of the genes with promoter HypoMarks/HyperMarks and ±3 kb flanking regions are illustrated as RNA read count per million mapped reads based on RNA-Seq data.

factor in the HypoMarks and HyperMarks of each cell type revealed that the cell type-specific HypoMarks are enriched in the binding sites for transcription factors involved in the regulation of the respective cell phenotype (Figure 4A). For example, the top two transcription factors in the hESC H1-specific HypoMarks were the well-known stem cell pluripotency factors POU5F1 and NANOG. In further support of this, the motif enrichment analysis based on the sequence of the cell type-specific HypoMarks/HyperMarks revealed many more interesting cell type-specific transcription factor associations, such as the enrichment of distinct FOXA2 in the pancreas HypoMarks, STAT3 in the esophagus HypoMarks, and CTCF in the H1 HyperMarks (Figure 4B, and Supplementary Table S6). The overlap of the enriched motifs between the HypoMarks and HyperMarks in the hESC H1 cells indicated the selective binding and regulation of transcription factors (Figure 4C). To verify this, we assessed the actual binding events of 50 transcription factors in the hESC H1 cells and constructed a transcription factor and MethyMark collaboration network for the hESC H1 cells (Figure 4D and Supplementary Figure S26A and B). We observed that the HypoMarks were bounded by the components of the transcription initiation complexes (for example, POLR2A), active transcription factors (for example, EP300, which might further induce the acetylation of H3K27) and hESC H1-specific active transcription factors (POU5F1 and NANOG) (Figure 4E). For example, the sub network constructed using the first neighbours of the transcription factors NANOG and POU5F1 revealed that the MethyMarks bounded by these two transcription factors are nearly all HypoMarks (Supplementary Figure S26C). The functional enrichment analysis confirmed four features of the genes associated with these HypoMarks, including targets of the transcription factors NANOG, POU5F1 and SOX2, overexpression in hESCs, functions associated with embryonic development and a relationship with abnormal developmental phenotypes (Supplementary Figure S26D). These findings suggested that the cell type-specific HypoMarks frequently co-localize with TFBSs and might facilitate the binding of transcription factors on DNA sequences in a cell type-specific manner.

The cell type-specific HypoMarks are associated with cell type-specific super-enhancers

Two features of the cell type-specific HypoMarks, H3K27ac and TFBSs, are also indicators of super-enhancers, which were recently identified as large clusters of transcriptional enhancers that drive the expression of the genes that define cell identity (21,22). Here, we identified the MethyMarks that overlapped with super-enhancers in the 21 cell types and observed a significant enrichment of the cell type-specific HypoMarks in super-enhancers of the same cell type (Figure 5A and Supplementary Table S7). Additional cell type-specific HypoMarks that only overlapped with super-enhancers from the same cell type were identified and treated as SuperHypoMarks (Table 2 and Supplementary Table S8). The cell type-specific SuperHypoMarks showed lower methylation and higher H3K27ac levels in specific cell types compared to other cell types (Supplementary Figure S27). In the same hESC cell type, H3K27ac was significantly negatively correlated with DNA methylation, and the SuperHypoMarks showed significantly higher H3K27ac levels and lower methylation than the other HypoMarks and HyperMarks (Figure 5B–D).

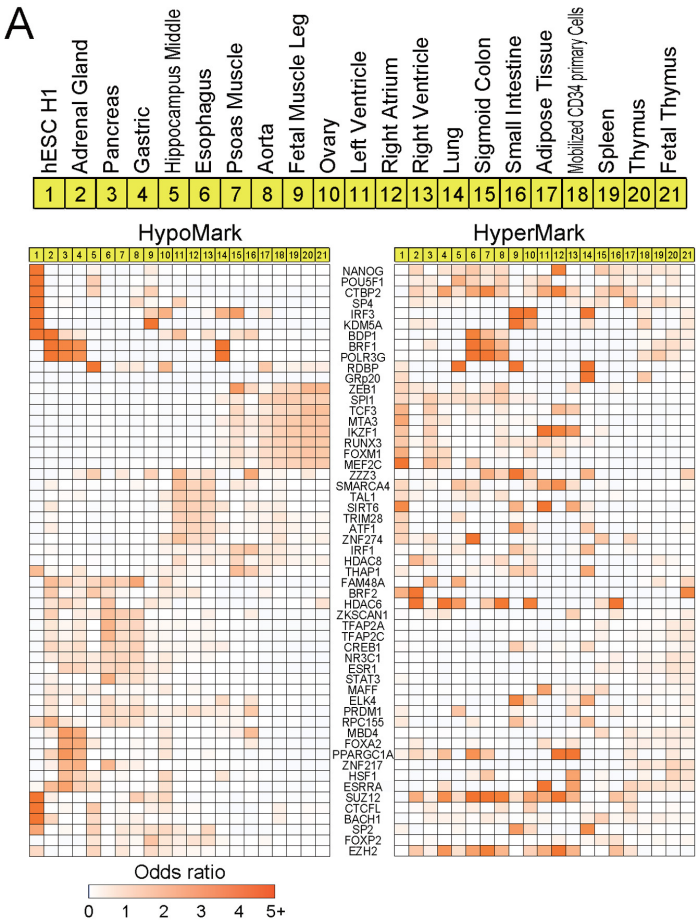
Table 2. The number of cell type-specific SuperHypoMarks and related genes

Cell-type	Number of SuperHypoMarks	Number of Related genes
hESC H1	175	71
Hippocampus middle	830	296
Adrenal gland	137	67
Esophagus	372	129
Ovary	225	57
Pancreas	188	67
Fetal muscle leg	365	110
Psoas muscle	268	90
Right atrium	36	19
Left ventricle	82	46
Right ventricle	4	3
Aorta	637	123
Gastric	364	128
Sigmoid colon	14	9
Small intestine	17	11
Lung	31	23
Adipose tissue	0	0
Spleen	174	86
Mobilized CD34 primary cells	104	51
Thymus	73	28
Fetal thymus	126	51

Note: The detailed information for the SuperHypoMarks in each cell type is listed in Supplementary Table S6.

An additional analysis revealed that the cell type-specific SuperHypoMarks were prone to be localized in close proximity to the genes involved in the regulation of the respective cellular states (Supplementary Table S8). For example, 58% (101/175) of the hESC H1-specific SuperHypoMarks were located in or nearby (with distance less than 2 kb) a unique set of 71 genes, including the two well-known hESC markers *POU5F1* and *NANOG* (Figure 5E). Intriguingly, the promoter region of the *POU5F1* gene showed distinct features, such as hESC H1-specific hypomethylation, super-enhancers, high H3K27ac and H3K4me3 levels, chromatin promoter/enhancer states and coactivator binding (Figure 5F and Supplementary Figure S28). The existence of hESC H1-specific SuperHypoMarks associated with transcription factor genes, the DNA methyltransferase *DNMT3B* and the histone methyltransferase *NSD1* (also known as *KMT3B*), and microRNAs suggest extensive regulation by the SuperHypoMarks (Supplementary Figure S28). In addition, the long noncoding RNAs, including *LINC00678,* associated with hESC H1-specific SuperHypoMarks might be potential novel markers for stem cells (Supplementary Figure S28).

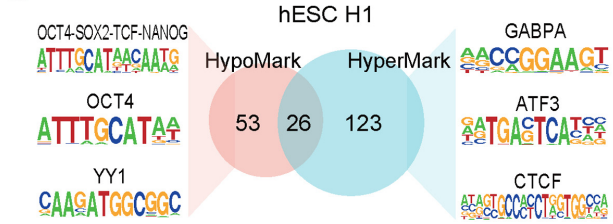
Finally, we asked whether the cell type-specific SuperHypoMarks were associated with lineage-specific mammalian pathways and phenotypes. To this end, we examined the mouse phenotypes that result from knockdowns of the mouse orthologues of the protein-coding genes associated with the cell type-specific SuperHypoMarks. The GREAT tool was used to determine the mouse phenotype term enrichment for the SuperHypoMarks from each of the rep-

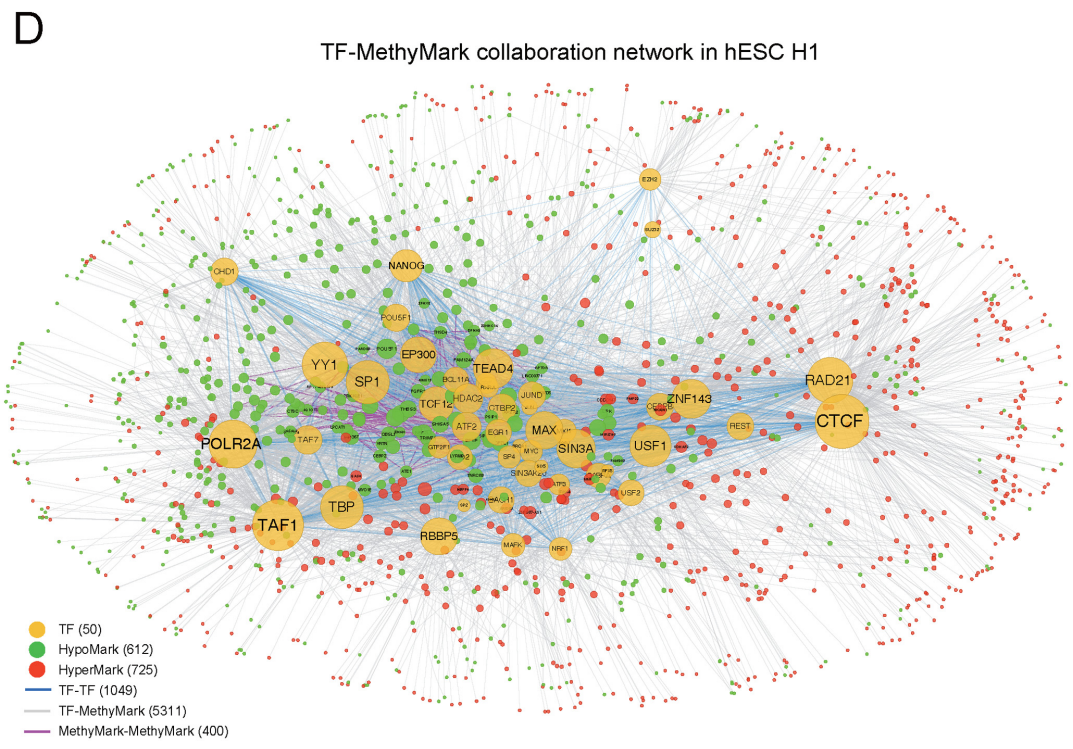


B

HypoMarks	Transcription factor	Motif	P-value
H1	OCT4-SOX2-TCF-NANOG		1.0×10^{-38}
Pancreas	FOXA2		1.0×10^{-187}
Esophagus	STAT3		1.0×10^{-25}
Fetal Thymus	TCF3		1.0×10^{-18}

HyperMarks	Transcription factor	Motif	P-value
H1	CTCF		1.0×10^{-3}
Pancreas	OCT4-SOX2-TCF-NANOG		1.0×10^{-8}
Esophagus	PRDM1		1.0×10^{-3}
Fetal Thymus	NP1		1.0×10^{-29}





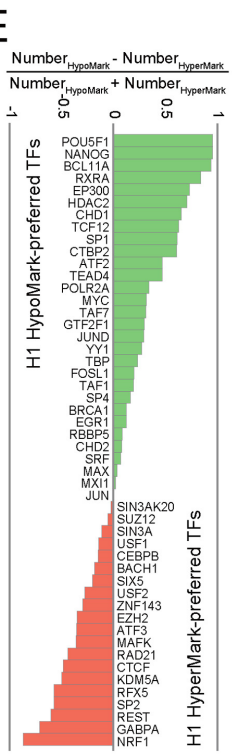


Figure 4. Cell type-specific HypoMarks collaborate with cell type-specific transcription factors. (**A**) Co-localization of the ENCODE uniform TFBSs with the HypoMarks and HyperMarks in 21 cell types. For each cell type, only the top four TFBSs according to the enrichment odds ratio are represented as a

resentative cell types, including stem cells (hESC H1), and all three germ layers, including the mesoderm (thymus), ectoderm (hippocampus middle) and endoderm (gastric) (Supplementary Table S9). As illustrated in Figure 5G, the top ten mouse phenotype categories were highly cell type-specific for all four representative cell types. For example, the knockdowns of the mouse orthologues of the protein-coding genes associated with the SuperHypoMarks in the hESC H1 cells may lead to abnormal development of multiple organs, such as the eye and cardiovascular, and nervous systems, while those in the hippocampus middle tissue may be related to abnormal oligodendrocyte morphology, myelination, synaptic transmission, etc. Meanwhile, the knockdowns of the mouse orthologues of the protein-coding genes associated with SuperHypoMarks in gastric tissue may cause abnormal gastric parietal cell morphology, and those in the Thymus may decrease the immunoglobulin levels. Taken together, these data have revealed that DNA methylation might play a key role in the control of the cell type specificity of the super-enhancers, which further regulate key identity genes in a cell type-specific manner.

DISCUSSION

DNA methylation plays important roles in gene regulation during cell development and differentiation, and aberrant methylation can cause multiple diseases, including cancer (10,51). The cell type-specific gene activity induced through cell type-specific methylation has been widely reported (13,31,34). A comprehensive map of the cell type-specific methylation marks is indispensable for the in-depth study of DNA methylation dynamics and regulatory mechanisms. The combination of bisulfite conversion and high-throughput sequencing offers the best quantitative method for studying DNA methylation at high resolution (4,10). The decreasing cost of sequencing promotes the profiling of DNA methylomes in various human cell lines and tissues, representing an unprecedented opportunity to identify the cell type-specific methylation marks and examine the features of the aberrations at the macro scale (3,11,28).

Here, we introduced a novel entropy-based framework to detect the cell type-specific methylation marks by integrating multiple methylomes from human cell lines and tissues. In this framework, Shannon entropy was optimized to quantify the methylation specificity across cell types for each CpG, and the distance-dependent methylation similarities between neighbouring CpGs was considered as the biological basis to merge CpGs into segments. Previous studies have shown the quantification of the differences in the methylation patterns in specific genome regions across multiple human samples and the identification of the differentially methylated regions using Shannon entropy to calculate the mean methylation of given regions (31). By optimizing the Shannon entropy, we showed that we can not only quantify the methylation specificity at single base precision, but we can also perform high-resolution genome segmentation through the integration of the BS-Seq methylomes from various cell types. The analysis of the methylation specificity across cell types revealed that several well-known cell type-specific regulatory elements exhibit high methylation specificity across human cell types, and additional analyses on the cell type specificity, genomic location and chromatin state of these segments classified the markers into different types, including HighSpe, InterSpe, LowSpe (UniHypo, UnipLow, UnipHigh and UniHyper), cell type-specific MethyMarks, Large MethyMarks and SuperHypoMarks (Figure 6A). The genome segmentation and functional regulatory element annotation based on the cell type-specificity of the DNA methylation patterns in the present study are important for the expansion of and contribution to the current knowledge of the functional DNA elements in the human genome.

These analyses revealed that the UniHyper segments comprise nearly half of the LowSpe segments, and most of these segments overlapped with repetitive elements, which accumulate throughout evolution and are usually silenced in the human genome (Figure 6A). Repetitive elements account for only 1.5% of the typical bacterial genomes and approximately 3% of the fly genome. In contrast, >50% of the human genome contains repeated sequences (52). The silencing of the repetitive elements in the human genome is essential for maintaining genomic stability. DNA methylation has been investigated for its roles in the control of genomic activity in most eukaryotic organisms, including plants, animals and fungi (53). In the present study, the conserved hypermethylation of repetitive elements across all studied human cell types, including pluripotent cells and adult tissues, also confirmed the critical role of DNA methylation in silencing repetitive elements. The aberrant hypomethylation of repetitive elements has been investigated in a variety of human diseases (54). For example, the global hypomethylation in tumors, which is a ubiquitous feature of carcinogenesis, primarily affects the hypomethylation of repetitive DNA sequences (51,55). In-depth studies on DNA methylation abnormalities in repetitive elements associated with human cancer will be helpful for understanding the mechanisms of carcinogenesis and proposing new treatments for cancer and other diseases.

In the 'desert' UniHyper segments, we also observed some 'oases', such as UniHypo segments (Figure 6A). These UniHypo segments were more similar to the CpG-dense regions localized in the promoter regions of known housekeeping genes, which are typically constitutive genes

←

color from white (low) to red (high). (**B**) Examples of known transcription factor motifs that are significantly enriched in the HypoMarks and HyperMarks. The *P*-value represents the significance of the enrichment of a motif in the HypoMarks/HyperMarks. (**C**) Overlap of the transcription factor motifs that were significantly enriched in the HypoMarks and HyperMarks in the hESC H1 cells. Three specifically enriched motifs in HypoMarks and HyperMarks are shown. (**D**) Transcription factor (TF) and MethyMark collaboration network in the hESC H1 cell line. The nodes and lines are indicated on the bottom left. The size of the transcription factor node represents the number of bound MethyMarks and the size of the MethyMark node represents the number of bound transcription factors. The width of the TF–TF line represents the number of MethyMarks that are targeted by two Transcription factors, while the MethyMark-MethyMark line represents the number of transcription factors binding to both MethyMarks. (**E**) The relative binding preference of 50 transcription factors to the hESC H1-specific HypoMarks and HyperMarks.

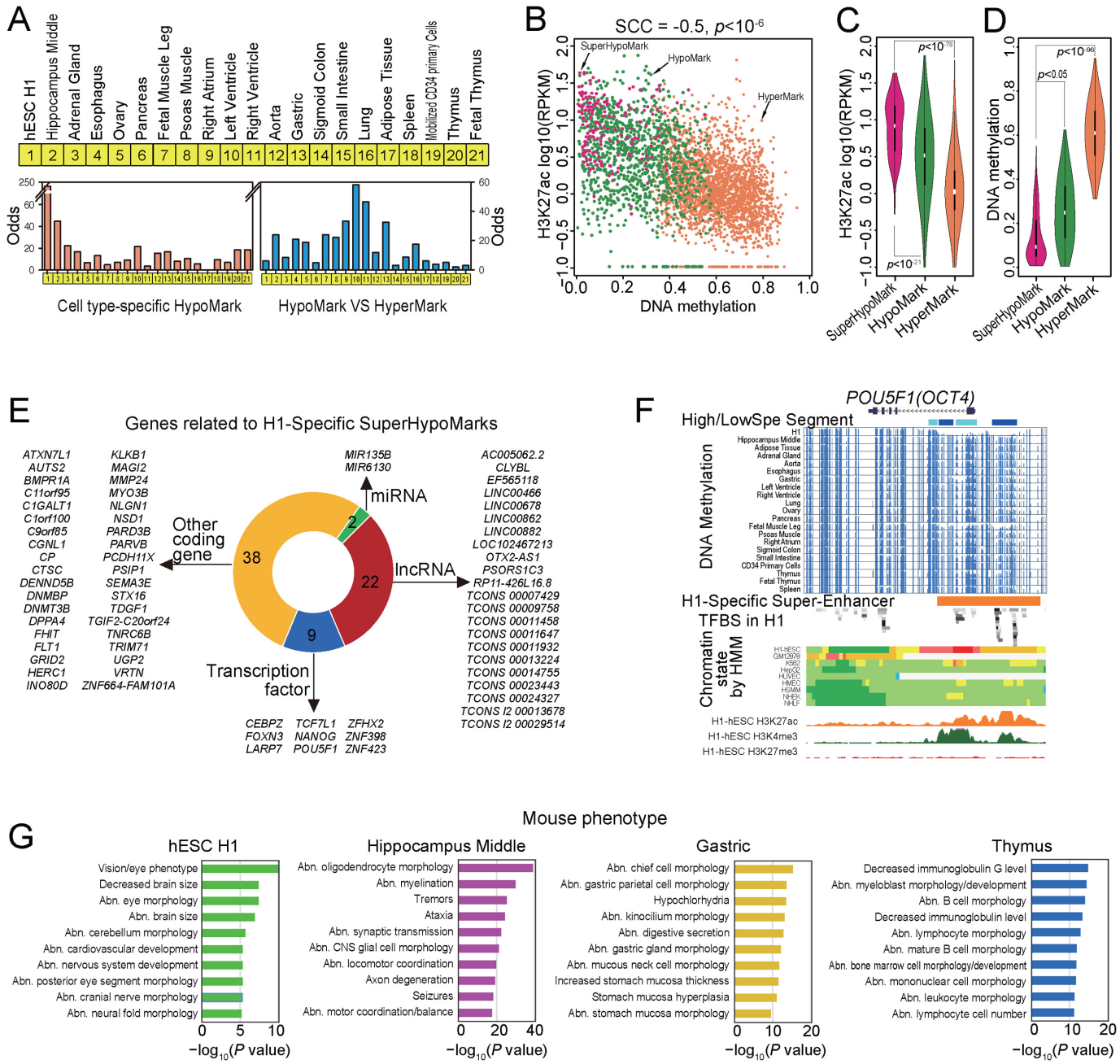
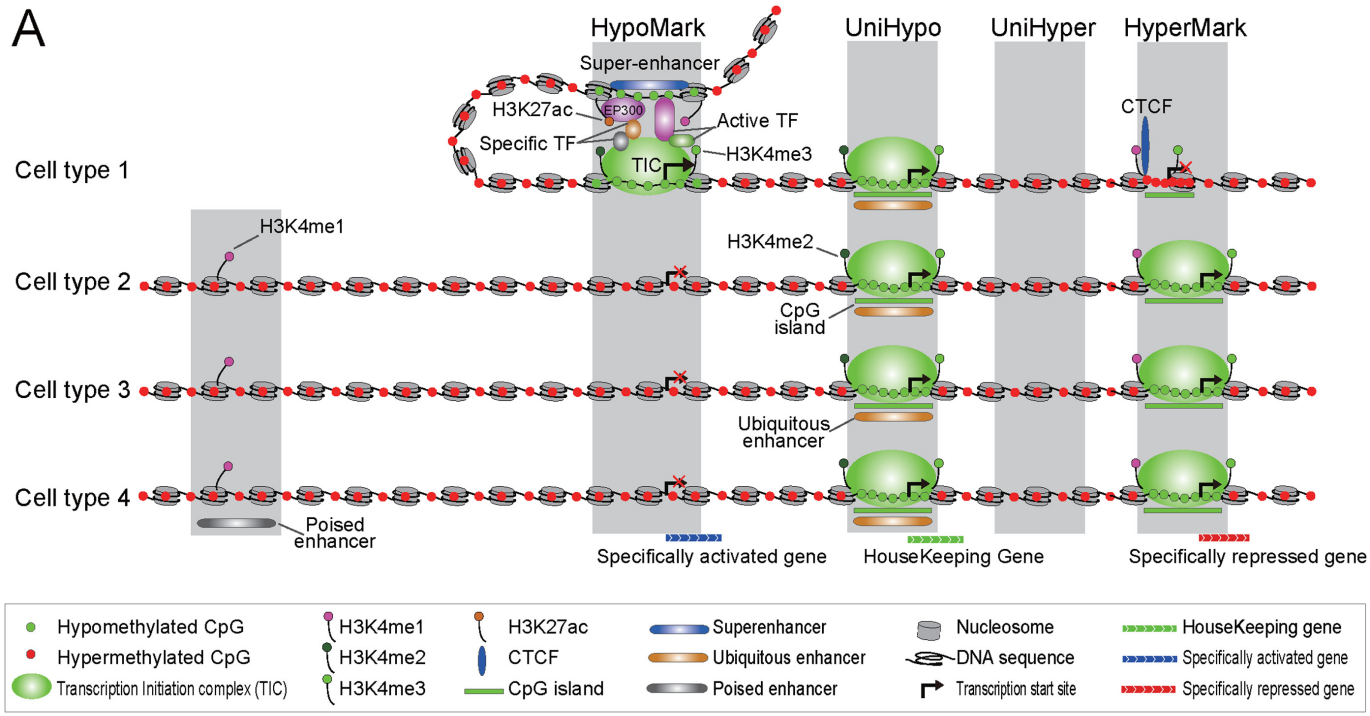


Figure 5. Cell type-specific SuperHypoMarks across 21 cell types. (**A**) Enrichment of the cell type-specific HypoMarks in the cell type-specific super-enhancers. For the super-enhancers in a specific cell type, the odds ratios of the HypoMarks of the same cell type compared to the HypoMarks of other cell types and the odds ratio of the HyperMarks of the same cell types are shown in the left and right panels, respectively. Detailed information is listed in Supplementary Table S7. (**B**) Scatter plot of the DNA methylation patterns and H3K27ac state of the H1 SuperHypoMarks, other HypoMarks and HyperMarks. SCC represents the Spearman's rank correlation coefficient between DNA methylation and H3K27ac, and *p* is the significance of the coefficient. (**C**) The box plot and kernel density plot of H3K27ac in the H1 SuperHypoMarks, other HypoMarks and HyperMarks. *P* shows the significance of the Wilcoxon rank sum test for the differences in the H3K27ac levels between the two groups. (**D**) The box plot and kernel density plot of DNA methylations in the H1 SuperHypoMarks, other HypoMarks and HyperMarks. (**E**) The genes associated with the hESC H1-specific SuperHypoMarks are shown. The detailed information about the epigenetic and wild-type expression of the genes in bold are given in Supplementary Figure S28. (**F**) An example of the hESC H1-specific SuperHypoMarks at the *POU5F1* locus. (**G**) The enrichment of the mouse phenotype terms associated with the cell type-specific SuperHypoMarks was calculated using the GREAT tool binomial approach. Detailed information is listed in Supplementary Table S9.



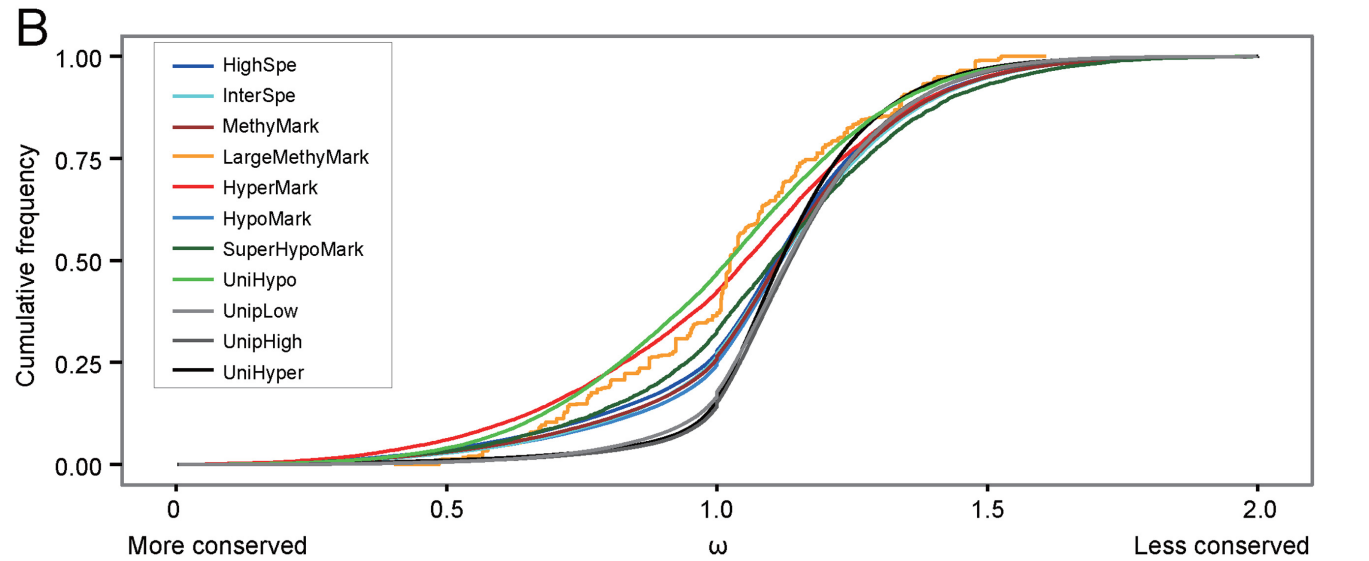


Figure 6. Human MethyMarks and sequence conservation. (**A**) Model for the human MethyMarks and regulatory elements. The model shows the methylation marks (including UniHypo, UniHyper and cell type-specific HypoMarks and HyperMarks), chromatin modifications and transcription factors detected in the regulatory regions of housekeeping genes and specifically activated or repressed genes. (**B**) The conservation levels (ω metric) for all of the types of methylation elements identified in the present study are shown as the cumulative distribution of sequence conservation for the HighSpe, InterSpe, UniHypo, UnipLow, UnipHigh and UniHyper segments, MethyMarks, Large MethyMarks, HyperMarks, HypoMarks and SuperHypoMarks across 29 mammalian species.

required for the maintenance of basic cellular functions (56). Studies have shown that most housekeeping genes have promoter CpG islands, which are typically Hypomethylated (57), and CpG islands are vertebrate genomic landmarks that often encompass the promoters of most housekeeping genes and are void of DNA methylation (58). We confirmed the stable hypomethylation of CpG islands across all of the studied human cell types, particularly those in housekeeping gene promoters. Additional analyses revealed a significant enrichment of the active chromatin modification H3K4me3 in the UniHypo segments, consistent with a previous finding that the non-methylated CpG-dense sequences recruit Cfp1 and establish H3K4me3 domains (59). In addition, these uniformly hypomethylated segments are

also enriched through H3K27ac and EP300, which are both markers of active enhancers, and these regions are typically identified as ubiquitous enhancers that overlap with CpG islands and are expressed in the majority of primary cells or tissues (20). In addition, a significant proportion of ubiquitous enhancers in the High/InterSpe segments indicated that the ubiquitous enhancers in a small number of tissues may undergo frequent DNA methylation changes in other tissues (Supplementary Figure S10A). It has been suggested that hypomethylation in the CpG islands of the promoters of housekeeping genes might be necessary for assembling active chromatin and ubiquitous enhancers, which are both required for ensuring the widespread expression of housekeeping genes in all cell types. These results indicate that the UniHypo segments comprise a small but distinct subset of methylation segments, which likely have specific regulatory functions in most human cell types. The high sequence conservation of the UniHypo segments across 29 mammalian species also demonstrates the importance of regulation by CpG island hypomethylation throughout the long history of evolution (Figure 6B).

The analysis of the MethyMark atlas revealed that the cell type-specific HypoMarks and HyperMarks exhibit distinct chromatin modifications and gene regulatory signatures (Figure 6A). As expected, the cell type specificity of DNA methylation has been strongly associated with well-known cell type-specific regulatory elements, including enhancers and super-enhancers. Both HypoMarks and HyperMarks are significantly enriched with the enhancer mark H3K4me1 and chromatin states corresponding to weak enhancers. These findings indicate that the cell type-specific MethyMarks might be indicators of enhancers. Unlike the enhancers with well-known chromatin marks, such as H3K4me1 and H3K27ac, cell type-specific suppressors are more difficult to identify on a large scale. Thus, the relationships between the DNA methylation marks and cell type-specific suppressors were not analyzed in this study. However, we found that some CpG island might play a regulatory role, similar to cell type-specific suppressors (Figure 6A). The enrichment of HyperMarks in CpG islands and gene promoter regions (Supplementary Figure S18A and B) and the low expression of the corresponding genes suggest that the cell type-specific hypermethylation of CpG islands might reflect the loss of enhancer activity, resulting in the selective inhibition of specific genes, which are not required in the specific cell types. In contrast, the cell type-specific HypoMarks showed a distinct enrichment of H3K27ac, an important indicator of active enhancers (49). This observation is consistent with previous studies showing hypomethylation in H3K27ac peaks and active enhancers (28,60). Recent studies have demonstrated that H3K27ac is a superior indicator for super-enhancer which represent a large cluster of transcriptional enhancers that drive the expression of the genes that define cell identity (21,22). In the present study, additional analyses revealed cell type-specific hypomethylation in cell type-specific super-enhancers. These results suggest that the cell type-specific HypoMarks and H3K27ac are important indicators of active enhancers, particularly cell type-specific super-enhancers.

The role of DNA hypomethylation in gene activation is currently unclear. Using the hormone-inducible glucocorticoid receptor as a model system (61), Wiench et al. observed that hypermethylation directly destabilizes interactions between the glucocorticoid receptor and DNA in vitro, and hormone-dependent demethylation at glucocorticoid receptor binding sites is associated with increased chromatin accessibility (23). In the present study, we observed the binding preference of components of the transcription initiation complex to HypoMarks. In addition, cell type-specific HypoMarks recruit transcription factors that are required in the corresponding cell type, which is consistent with a recent finding based on the DNA methylation data of 54 normal cell lines that were profiled using reduced representation bisulfite sequencing (62). For example, the hESC H1-specific HypoMarks are uniquely bound by the pluripotency transcription factors NANOG and POU5F1, consistent with a previous finding that the ESC master transcription factors Pou5f1, Sox2, and Nanog form super-enhancers at most genes to control the pluripotent state (21). Thus, we propose that cell type-specific hypomethylation might play a role in recruiting cell type-specific transcription factors and assembling the transcription machinery at super-enhancers, which further promote distinct gene expression profiles for the characterization of cellular phenotypes (Figure 6A). This hypothesis has been confirmed by recent studies of the transcription factor binding dynamics during human ES cell differentiation (30), and the cell line-specific epigenetic modification models for transcription factor binding predictions (63). More direct evidence is obtained from a recently published study, which revealed the DNA hypomethylation-mediated regulation of the cell identity Myf5 super-enhancer in the establishment of the skeletal muscle lineage (64). These results indicated that DNA methylation might play a key role in the control of the cell type specificity of the super-enhancers, although this may need to be confirmed by experiments in additional cell types. Further research is required to determine the mechanism, which may generate important knowledge on new ways to manipulate the cell fate potential of stem cells and mature adult cells. The high conversion level of the Super-HypoMarks suggested that regulation through cell type-specific hypomethylation might be widespread in mammals (Figure 6B). Notably, regulation through hypomethylation might represent a double-edged sword. Aberrant hypomethylation might assemble super-enhancers that are not normally programmed in a specific cell type, resulting in disease phenotypes, such as cancers (65,66). This idea suggests that the hypotheses regarding the role of DNA methylation and genes in many diseases might be based on the knowledge of HypoMarks, particularly SuperHypoMarks. Further mining of disease-specific SuperHypoMarks and examinations of the underlying mechanisms should be useful for the diagnosis, prognosis and treatment of complex diseases.

We also identified 5 406 large methylation segments with length of at least 3.5 kb, which were used to identify long hypomethylated genomic regions (12). Among these large segments, 169 segments were uniformly hypomethylated in all of the studied cell types, providing additional support for the conservation of long hypomethylated regions, such as DNA methylation valleys and canyons, across normal cell types (11,12). Interestingly, 65 large methylation segments

were identified as cell type-specific HyperMarks, which are hypermethylated in a minority of cell types and hypomethylated in the majority of cell types. This finding suggests that large uniformly hypomethylated regions might not be conserved when more diverse cell types are considered. Importantly, we observed that large cell type-specific MethyMarks are associated with cell type-specific identity genes and imprinted genes, which are highly conversed across mammals (Figure 6B). Thus, we concluded that the large cell type-specific MethyMarks are key elements of gene regulatory domains that are associated with cell type-specific phenotypes.

Sequencing-based DNA methylomes and a novel integrative entropy-based tool SMART were used to map a comprehensive atlas of the cell type-specific methylation marks across multiple human cell lines and tissues. The findings underscore the importance of DNA methylation as a stable marker of regulatory elements for cell identity. SMART, combined other BS-Seq data analysis tools, such as CpGMPs (29), RnBeads (30), QDMR (31) and DSS-single (67), should be used to fine map the methylation marks in more specific subpopulations, including different subtypes of cells with normal or disease phenotypes. An additional integrative analysis of the methylation marks in the reference human epigenomes (36,68) should considerably advance our understanding of the cell type-specific methylation machinery in the regulation of transcriptional activity and modulation of the cellular phenotypes. Particularly, the mining of pluripotency-associated methylation marks would enhance the applicability of DNA methylation as a marker for embryonic stem cells or induced pluripotent stem cells (69,70). The in-depth study of more disease-specific methylation marks should be sufficient to diagnose disease by profiling only a representative subset of CpG sites in the well-defined marks via gel-based or array-based technologies (71). Overall, we hope that the framework and atlas proposed in the present study are valuable for exploring the association of DNA methylation with regulatory dynamics and cell identity elements in additional phenotypes and species.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the NIH Roadmap Epigenomics and ENCODE consortia for generating and sharing the data used in this paper. We thank Hui Liu, Yihan Wang for discussions related to this work. We thank Min Shao and Yunzhen Wei for advice on the algorithm design.

FUNDING

National Natural Science Foundation of China [61403112, 31371334, 81573021, 61402139 and 31371478]; Natural Scientific Research Fund of Heilongjiang Provincial [ZD2015003]. Funding for open access charge: National Natural Science Foundation of China [61403112].
Conflict of interest statement. None declared.

REFERENCES

1. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
2. Holliday,R. and Pugh,J.E. (1975) DNA modification mechanisms and gene activity during development. *Science*, **187**, 226–232.
3. Ziller,M.J., Gu,H., Muller,F., Donaghey,J., Tsai,L.T., Kohlbacher,O., De Jager,P.L., Rosen,E.D., Bennett,D.A., Bernstein,B.E. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
4. Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
5. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
6. Irizarry,R.A., Ladd-Acosta,C., Wen,B., Wu,Z., Montano,C., Onyango,P., Cui,H., Gabo,K., Rongione,M., Webster,M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.
7. Rakyan,V.K., Down,T.A., Thorne,N.P., Flicek,P., Kulesha,E., Graf,S., Tomazou,E.M., Backdahl,L., Johnson,N., Herberth,M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, **18**, 1518–1529.
8. Lokk,K., Modhukur,V., Rajashekar,B., Martens,K., Magi,R., Kolde,R., Kolt Ina,M., Nilsson,T.K., Vilo,J., Salumets,A. *et al.* (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.*, **15**, R54.
9. Court,F., Tayama,C., Romanelli,V., Martin-Trujillo,A., Iglesias-Platas,I., Okamura,K., Sugahara,N., Simon,C., Moore,H., Harness,J.V. *et al.* (2014) Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res.*, **24**, 554–569.
10. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
11. Xie,W., Schultz,M.D., Lister,R., Hou,Z., Rajagopal,N., Ray,P., Whitaker,J.W., Tian,S., Hawkins,R.D., Leung,D. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134–1148.
12. Jeong,M., Sun,D., Luo,M., Huang,Y., Challen,G.A., Rodriguez,B., Zhang,X., Chavez,L., Wang,H., Hannah,R. *et al.* (2014) Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat. Genet.*, **46**, 17–23.
13. Song,F., Smith,J.F., Kimura,M.T., Morrow,A.D., Matsuyama,T., Nagase,H. and Held,W.A. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3336–3341.
14. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.
15. Zhang,B., Zhou,Y., Lin,N., Lowdon,R.F., Hong,C., Nagarajan,R.P., Cheng,J.B., Li,D., Stevens,M., Lee,H.J. *et al.* (2013) Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res.*, **23**, 1522–1540.
16. Hu,S., Wan,J., Su,Y., Song,Q., Zeng,Y., Nguyen,H.N., Shin,J., Cox,E., Rho,H.S., Woodard,C. *et al.* (2013) DNA methylation presents distinct binding sites for human transcription factors. *eLife*, **2**, e00726.
17. Medvedeva,Y.A., Khamis,A.M., Kulakovskiy,I.V., Ba-Alawi,W., Bhuyan,M.S., Kawaji,H., Lassmann,T., Harbers,M., Forrest,A.R., Bajic,V.B. *et al.* (2014) Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, **15**, 119.
18. Wang,H., Maurano,M.T., Qu,H., Varley,K.E., Gertz,J., Pauli,F., Lee,K., Canfield,T., Weaver,M., Sandstrom,R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.
19. Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
20. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T.

et al. (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
21. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
22. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-Andre,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
23. Wiench,M., John,S., Baek,S., Johnson,T.A., Sung,M.H., Escobar,T., Simmons,C.A., Pearce,K.H., Biddie,S.C., Sabo,P.J. *et al.* (2011) DNA methylation status predicts cell type-specific enhancer activity. *EMBO J.*, **30**, 3028–3039.
24. Ko,Y.A., Mohtat,D., Suzuki,M., Park,A.S., Izquierdo,M.C., Han,S.Y., Kang,H.M., Si,H., Hostetter,T., Pullman,J.M. *et al.* (2013) Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. *Genome Biol.*, **14**, R108.
25. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
26. Rivera,C.M. and Ren,B. (2013) Mapping human epigenomes. *Cell*, **155**, 39–55.
27. Lister,R., Pelizzola,M., Kida,Y.S., Hawkins,R.D., Nery,J.R., Hon,G., Antosiewicz-Bourget,J., O'Malley,R., Castanon,R., Klugman,S. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.
28. Hon,G.C., Rajagopal,N., Shen,Y., McCleary,D.F., Yue,F., Dang,M.D. and Ren,B. (2013) Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat .Genet.*, **45**, 1198–1206.
29. Su,J., Yan,H., Wei,Y., Liu,H., Wang,F., Lv,J., Wu,Q. and Zhang,Y. (2013) CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res.*, **41**, e4.
30. Tsankov,A.M., Gu,H., Akopian,V., Ziller,M.J., Donaghey,J., Amit,I., Gnirke,A. and Meissner,A. (2015) Transcription factor binding dynamics during human ES cell differentiation. *Nature*, **518**, 344–349.
31. Zhang,Y., Liu,H., Lv,J., Xiao,X., Zhu,J., Liu,X., Su,J., Li,X., Wu,Q., Wang,F. *et al.* (2011) QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.*, **39**, e58.
32. Genereux,D.P., Johnson,W.C., Burden,A.F., Stoger,R. and Laird,C.D. (2008) Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Res.*, **36**, e150.
33. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
34. Liu,H., Chen,Y., Lv,J., Zhu,R., Su,J., Liu,X., Zhang,Y. and Wu,Q. (2013) Quantitative epigenetic co-variation in CpG islands and co-regulation of developmental genes. *Sci. Rep.*, **3**, 2576.
35. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
36. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
37. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
38. Shen,L., Shao,N., Liu,X. and Nestler,E. (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.
39. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
40. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
41. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
42. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
43. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
44. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14863–14868.
45. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
46. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
47. Garber,M., Guttman,M., Clamp,M., Zody,M.C., Friedman,N. and Xie,X. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, **25**, i54–i62.
48. Wei,Y., Su,J., Liu,H., Lv,J., Wang,F., Yan,H., Wen,Y., Liu,H., Wu,Q. and Zhang,Y. (2014) MetaImprint: an information repository of mammalian imprinted genes. *Development*, **141**, 2516–2523.
49. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
50. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
51. Esteller,M. (2008) Epigenetics in cancer. *N. Engl. J. Med.*, **358**, 1148–1159.
52. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
53. Zemach,A., McDaniel,I.E., Silva,P. and Zilberman,D. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, **328**, 916–919.
54. Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.
55. Ehrlich,M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239–259.
56. Zhu,J., He,F., Hu,S. and Yu,J. (2008) On the nature of human housekeeping genes. *Trends Genet.*, **24**, 481–484.
57. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
58. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
59. Thomson,J.P., Skene,P.J., Selfridge,J., Clouaire,T., Guy,J., Webb,S., Kerr,A.R., Deaton,A., Andrews,R., James,K.D. *et al.* (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, **464**, 1082–1086.
60. Bell,J.T., Pai,A.A., Pickrell,J.K., Gaffney,D.J., Pique-Regi,R., Degner,J.F., Gilad,Y. and Pritchard,J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.
61. Thomassin,H., Flavin,M., Espinas,M.L. and Grange,T. (2001) Glucocorticoid-induced DNA demethylation and gene memory during development. *EMBO J.*, **20**, 1974–1983.
62. Yang,X., Shao,X., Gao,L. and Zhang,S. (2015) Systematic DNA methylation analysis of multiple cell lines reveals common and specific patterns within and across tissues of origin. *Hum. Mol. Genet.*, **24**, 4374–4384.
63. Liu,L., Jin,G. and Zhou,X. (2015) Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res.*, **43**, 3873–3885.
64. Carrio,E., Diez-Villanueva,A., Lois,S., Mallona,I., Cases,I., Forn,M., Peinado,M.A. and Suelves,M. (2015) Deconstruction of DNA methylation patterns during myogenesis reveals specific epigenetic

events in the establishment of the skeletal muscle lineage. *Stem Cells*, **33**, 2025–2036.
65. Affer,M., Chesi,M., Chen,W.D., Keats,J.J., Demchenko,Y.N., Tamizhmani,K., Garbitt,V.M., Riggs,D.L., Brents,L.A., Roschke,A.V. *et al.* (2014) Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. *Leukemia*, **28**, 1725–1735.
66. Loven,J., Hoke,H.A., Lin,C.Y., Lau,A., Orlando,D.A., Vakoc,C.R., Bradner,J.E., Lee,T.I. and Young,R.A. (2013) Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, **153**, 320–334.
67. Wu,H., Xu,T., Feng,H., Chen,L., Li,B., Yao,B., Qin,Z., Jin,P. and Conneely,K.N. (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1715.
68. Schultz,M.D., He,Y., Whitaker,J.W., Hariharan,M., Mukamel,E.A., Leung,D., Rajagopal,N., Nery,J.R., Urich,M.A., Chen,H. *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.
69. Cacchiarelli,D., Trapnell,C., Ziller,M.J., Soumillon,M., Cesana,M., Karnik,R., Donaghey,J., Smith,Z.D., Ratanasirintrawoot,S., Zhang,X. *et al.* (2015) Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. *Cell*, **162**, 412–424.
70. Farlik,M., Sheffield,N.C., Nuzzo,A., Datlinger,P., Schonegger,A., Klughammer,J. and Bock,C. (2015) Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.*, **10**, 1386–1397.
71. Lv,J., Liu,H., Su,J., Wu,X., Li,B., Xiao,X., Wang,F., Wu,Q. and Zhang,Y. (2012) DiseaseMeth: a human disease methylation database. *Nucleic Acids Res.*, **40**, D1030–D1035.