Spring 5-15-2020

Developing Tools for Identifying Tissue-Specific Epigenetic Marks and Predicting DNA hydroxy/methylation

Yu He

Washington University in St. Louis

Recommended Citation

He, Yu, "Developing Tools for Identifying Tissue-Specific Epigenetic Marks and Predicting DNA hydroxy/methylation" (2020). *Arts & Sciences Electronic Theses and Dissertations*. 2197.
https://openscholarship.wustl.edu/art_sci_etds/2197

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Ting Wang, Chair
Michael Brent
Jeremy Buhler
Harrison Gabel
Nan Lin

Developing Tools for Identifying Tissue-Specific Epigenetic Marks and Predicting DNA hydroxy/methylation

by
Yu He

A dissertation presented to
The Graduate School of
Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2020
St. Louis, MO

Table of Contents

List of Figures

Chapter 2

Supplementary Figures

Chapter 3

<u>Supplementary Figures</u>

List of Tables

Chapter 2

Chapter 3

Acknowledgments

I would not achieve PhD without my advisor, Dr. Ting Wang. In the past 5 years, there were so many times where I felt struggled by the research and disturbed by life. Ting never gave up on me and always patiently studied the problems I had and supported me with great encouragement. I would forever remembere how Ting almost rewrote my first paper, which astonished me for the rigorism he showed. Ting is a great model example for me as he pays so much enthusiasm towards his work all the time. Ting, thank you so much.

I want to thank my wife, Jinjin Qin, for the greatest support on me during the graduate school. Without her, I would not be able to finish my PhD. She not only helped me take care of my son while I was preparing qualifying exam, writing papers and preparing for thesis defense, but also consistently encouraged me when I was struggled by work and life. Thank you so much, Jinjin.

I want to thank the whole Wang lab. Everyone in the lab is so kind and generous to each other. The environment is like home to me. I want to thank Josh Jang, who contributed greatly to my second paper. Without him, I won’t be able to finish my thesis. Thank for your encouragement and support when I struggled, Josh. I want to thank Bo Zhang for mentoring me and giving me great ideas for EpiCompare paper. I want to thank Deepak Purushotham for installing Shiny framework for EpiCompare paper. I am very sorry that I forgot to put Bo and Deepak’s names in the paper.

I want to thank my thesis committee, Dr. Jeremy Buhler, Dr. Michael Brent, Dr. Harrison Gabel, Dr. Nan Lin for their valuable time on my thesis and advising me on critical questions. I want to thank the WashU community for providing the perfect work environment.

Yu He

Washington University in St. Louis

May 2020

Dedicated to my wife, Jinjin Qin.

ABSTRACT OF THE DISSERTATION

Developing Tools for Identifying Tissue-Specific Epigenetic Marks and Predicting DNA hydroxy/methylation

by

Yu He

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2020

Professor Ting Wang, Ph.D. Chair

A single genome can derive phenotypically unique cell types through various epigenetic modifications that instruct specific gene expression patterns. Histone modifications, DNA methylation, and DNA hydroxymetylation are the most common epigenetic modifications. To understand the mechanisms how these epigenetic modifications regulate gene expression, one often needs to map these marks genome-wide through profiling methods. Firstly, for histone modifications, Roadmap Epigenomics Consortium generated The Human Reference Epigenome Map, containing thousands of genome-wide histone modification datasets that describe epigenomes of a variety of different human tissue and cell types. This map has allowed investigators to obtain a much deeper and more comprehensive view of our regulatory genome, e.g. defining regulatory elements including all promoters and enhancers for a given tissue or cell type. An outstanding task is to combine and compare different epigenomes in order to identify regions with epigenomic features specific to certain types of tissues or cells, e.g. lineage-specific regulatory elements. Currently available tools do not directly address this question. This need motivated us to develop a tool that allows investigators to easily identify regions with epigenetic features unique to specific epigenomes that they choose, making detection of common regulatory elements and/or cell type- specific regulatory elements an interactive and dynamic experience. An online tool EpiCompare was developed to assist investigators in exploring the specificity of epigenomic features across selected tissue and cell types. Investigators can design their test by choosing different combinations of epigenomes, and choosing different classification algorithms provided by our tool. EpiCompare will then identify regions with specified epigenomic features, and provide a quality assessment of the predictions. Investigators can interact with EpiCompare by investigating Roadmap Epigenomics data, or uploading their own data for comparison. We demonstrated that by using specific combinations of epigenomes we can detect developmental lineage-specific enhancers. Secondly, for DNA methylation and hydroxymethylation, generating high resolution methylomes and hydroxymethylomes is a significant barrier for individual laboratories, therefore so far only a few cell types have deeply sequenced hydroxymethylomes at single-base resolution. This potential cost-barrier problem engendered a need for cost-effective, but high-resolution 5hmC mapping technology. Current enrichment-based technologies provide cheap, but low-resolution and relative enrichment of 5hmC levels while single base-resolution methods can be prohibitively expensive to scale up to large experiments. To address this problem, we develop a deep learning-based method "DeepH&M", which integrates enrichment and restriction enzyme sequencing methods to simultaneously estimate absolute hydroxymethylation and methylation levels at single CpG resolution. Using 7-week-old mouse cerebellum data for training DeepH&M model, we demonstrate that the 5hmC and 5mC levels predicted by DeepH&M were in high concordance with whole genome bisulfite- based approaches. The DeepH&M model can be applied to 7-week old frontal cortex and 79-week cerebellum revealing the robust generalizability of this method to other tissues from various biological time points.

Chapter 1: Introduction

One of the fundamental mysteries in biology is the generation of diverse cell types. Nearly every cell type in an organism shares the same genomic material but each cell type has different gene expression pattern and exerts different function and phenotype. Enormous amounts of evidence indicate that the epigenome instructs the gene expression program of different cell types with the genome. Histone modifications, DNA methylation, and DNA hydroxymetylation are the most common epigenetic modifications.

To understand the mechanisms how epigenetic modifications regulate gene expression, one often needs to map these marks genome-wide through profiling methods. Firstly, for histone modifications, Roadmap Epigenomics Consortium generated The Human Reference Epigenome Map, containing thousands of genome-wide histone modification datasets that describe epigenomes of a variety of different human tissue and cell types. This map has allowed investigators to obtain a much deeper and more comprehensive view of our regulatory genome, e.g. defining regulatory elements including all promoters and enhancers for a given tissue or cell type. I developed an online tool “EpiCompare”, which compares different epigenomes in order to identify regions with epigenomic features specific to certain types of tissues or cells, e.g. lineage-specific regulatory elements. Secondly, for DNA methylation and hydroxymethylation, generating high resolution methylomes and hydroxymethylomes is a significant barrier for individual laboratories. Current enrichment-based technologies provide cheap, but low-resolution and relative enrichment of 5hmC levels while single base-resolution methods can be prohibitively expensive to scale up to large experiments. I developed a deep learning-based method “DeepH&M”, which integrates enrichment and restriction enzyme sequencing methods to simultaneously estimate absolute hydroxymethylation and methylation levels at single CpG resolution.

1.1 The Epigenome

The epigenome refers to all chemical modifications of the chromatin including posttranslational modifications on histone proteins and DNA methylation and hydroxymethylation. Chromosomes are composed of nucleosome units, which are packed by DNA wrapping histone octamers, including H2A, H2B, H3, H4 and their variants. The tails and globular domains of histone proteins are subject to diverse posttranslational modifications. These histone modifications can directly affect chromatin accessibility by altering the net charge of histone proteins and thus changing chromatin structure, or serving as substrate for chromatin binding proteins, such as chromatin modifying complexes[1]. DNA methylation in the human genome primarily occurs at cytosine's fifth carbon (5mC) and cytosine methylation is largely restricted to CpG dinucleotides. DNA hydroxymethylation is an oxidative product of 5mC in which the hydrogen atom at the C5-position in cytosine is replaced by a hydroxymethyl group (5hmC). The methylation and hydroxymethylation of cytosine can affect the transcription of genes by impeding the binding of transcription factors or recruiting proteins bound to methylated cytosine.

1.1.1 Histone modifications

Over 130 posttranslational modifications on histone proteins and over 700 distinct histone isoforms have been identified so far[2,3]. The identified histone modifications include methylation, acetylation, propionylation, butyrylation, formylation, phosphorylation, ubiquitylation, sumoylation, citrullination, proline isomerization, ADP ribosylation, hydroxylation, and crotonylation. Initially, chromatin immunoprecipitation followed by DNA microarray (ChIP-

chip) was used to map genome-wide binding profile of chromosomal proteins[4]. With the advent of next generation sequencing technology, chromatin immunoprecipitation followed by sequencing (ChIP-seq) became the main method for mapping genome-wide binding profile[5]. With the characterization of histone modifications on a genome-wide scale using these technologies, many individual histone modifications have been found to be associated with specific functional elements[5–9]. First, H3K4me3 was found to localize to promoters and associated with transcription initiation, and H3K36me3 was detected at gene body and associated with transcription elongation[10,11]. Enhancers were at first thought to have similar histone marks as promoters, but later were found to enrich for H3K4me1 instead of H3K4me3[8]. However, the H3K4me1 or H3K4me3 mark alone is not sufficient to activate gene expression. H3K27ac is shown to distinguish active enhancers and promoters from inactive enhancers and promoters[6]. As for repressed chromatin, two distinct types have been identified. One is H3K9me3-marked heterochromatin, which is concentrated in pericentromeric regions; the other is H3K27me3-marked regions, which repress cell-type specific genes[12].

Although individual chromatin marks are associated with various functional elements, different chromatin marks can occur at the same locations, as confirmed by sequential ChIP-seq experiments[13]. One example is the identification of bivalent promoter state marked by both H3K4me3 and H3K27me3. Bivalent promoter states silence developmental genes in embryonic stem cells (ES) and keep them poised for activation in differentiated cells[13]. Therefore, it is possible that different chromatin marks combine together to encode function. This is the histone code hypothesis that was proposed over 20 years ago after the identification of large number of histone modifications[14].

1.1.2 Chromatin states

The combinations of chromatin marks that are biologically meaningful and recurrent throughout the genome are called chromatin states. Various computational algorithms based on Hidden Markov Models (HMM), clustering methods and others have been developed to define chromatin states. They integrate a collection of chromatin mark datasets and generate non-overlapping segmentations of the whole genome and assign labels to each segment. The labels generated are then interpreted as biologically meaningful chromatin states based on enrichment of known functional annotations, sequence motifs, and some specific experimentally observed characteristics. These identified chromatin states include promoter state, transcription state, enhancer state, repressed state, and quiescent state. Each chromatin state is enriched for distinct combinations of chromatin marks[15–19]. For example, active promoter state is enriched for H3K4me3, H3K4me2, H3K9ac, and H3K27ac, and bivalent promoter state is enriched for H3K4me3, H3K4me2 and H3K27me3[15]. Active enhancer state is enriched for H3K4me1 and H3K27ac, and bivalent enhancer state is enriched for H3K4me1 and H3K27me3. All promoter states have lower DNA methylation and all enhancer states have intermediate DNA methylation[20]. Transcribed state is enriched for H3K36me3, H3K79me3, H3K79me2, H3K79me1, H3K27me1, H2BK5me1, H4K20me1, and high DNA methylation. Heterochromatin is enriched for H3K9me3 and H3K9me2 and high DNA methylation.

The identification of chromatin states generates systematic annotations of the genome in multiple species, including human, mouse, fly, worm, yeast and plants[12,16,21–24]. The annotations include promoter, enhancer, insulator, transcribed, repressed and quiescent states. These annotations agree largely with genomic annotations. For example, in the human genome, promoter states identified overlap with over 90% of RefSeq TSS and transcribed states overlap with over 90% of RefSeq genes[16,19]. Around 5% of the genome is in enhancer state[15]. Luciferase reporter assays have shown that strong enhancer states have much stronger reporter activity than weak enhancer states. Clustering of enhancer states among multiple cell types identified cell-type specific enhancer state and they are enriched for biological processes and transcription factor binding sites in that cell type[15,20]. Studies have also shown that strong enhancer states significantly overlap with disease-causing variants and many of them are enriched in enhancers specific to disease-related cell types[15,18–20,25]. For example, Farh et. al. found that 60% of candidate causal variants from 21 autoimmune diseases mapped to immune-cell specific enhancers. 10%-20% of these causal variants altered transcription factor binding sites and thus altered gene expression.

Over 15 years have passed since the completion of the human genome sequencing, but we are still not completely clear about the function of the entire human genome. Only around 1.5% of the human genome is protein-coding[26]. In contrast, over 98% of the genome is non-coding, including regulatory elements such as enhancers, promoters, silencers, and insulators. Understanding the non-coding genome is important because regulatory elements can control the transcription of genes. It is also important for understanding the contribution of regulatory elements in diseases. Over 90% of GWAS hits are located in non-coding regions, and yet hard to interpret because of the lack of annotations in non-coding regions[27]. Of these regulatory elements, enhancers are the key players in regulating the spatial and temporal gene expression. Dysfunction of enhancers are often linked to disease. Plenty of evidence have shown that enhancers overlap significantly with disease-causing SNPs[15,18,19]. Hence, a comprehensive list of enhancer locations could have significant diagnostic potential. Identifying enhancers has been challenging because they are usually far from promoters and do not have common sequence signatures. Furthermore, over 60% of enhancers are cell-type specific, making identification of enhancers in each cell type necessary[28,29].

1.1.3 CpG DNA methylation and hydroxymethylation

DNA methylation has been shown to play a vital role in gene regulation, genomic imprinting, X-chromosome activation, the repression of transposable elements, etc[30–33]. Generally, high level of DNA methylation is associated with repression. The methylation of cytosine can affect the transcription of genes by impeding the binding of transcription factors or recruiting proteins bound to methylated cytosine. The pattern of 5mC is globally similar between different cell types but changes at specific loci can affect cell fate decisions. Although 60%-80% of 28 million CpG dinucleotides in human genome are methylated, less than 10% of CpG occur in CG-dense regions called CpG islands that are largely unmethylated[34]. These unmethylated CpG islands are enriched at promoters of housekeeping genes and developmentally regulated genes. Recently, tissue-specific hypo-methylated regions have been found to be enriched in enhancers and are associated with activation of targeted genes[35,36].

Although DNA methylation has been known for decades, how DNA methylation is actively removed besides passive loss of DNA methylation during replication was not discovered until recently. In 2009, two breakthrough paper discovered high abundance of 5-hydroxymethylation (5hmC), the oxidative product of 5mC, in mouse embryonic stem cells and Purkinje neurons[37,38]. A family of ten-eleven translocation (TET) proteins including TET1, TET2, TET3 were found to oxidize 5mC to 5hmC. Subsequently, TET enzymes were shown to oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) stepwise[39]. 5fC and 5caC can be further excised by DNA repairing enzyme thymine DNA glycosylase and restored to unmodified cytosine through base-excision repairing process, thus completing the process of active demethylation[40,41].

Since the discovery of active DNA demethylation process, many quantification and sequencing methods have been developed to study the global level and genome-wide distribution of 5hmC, 5fC, and 5caC[42]. 5hmC is abundant and as high as 40% of 5mC level in Purkinje neurons[37], ~ 5% of 5mC level in embryonic stem cells (ESCs)[43], and is low (less than 1% of 5mC level) in other cell types[44]. Compared to 5hmC, 5fC and 5caC are extremely low (hundred times lower than 5hmC) in cells[39]. In this dissertation, I mainly focus on 5mC and 5hmC. Genome-wide profiling of 5hmC has found that 5hmC is not just an intermediate of the active DNA demethylation process, but also a stable epigenetic mark involved in gene regulation. 5hmC is enriched in promoters, gene bodies and enhancers[44–46]. 5hmC level in promoters and gene bodies positively correlates with gene expression[45,47]. 5hmC level in enhancers positively correlates with active enhancer marks such as H3K4me1 and H3K27ac and the identification of cell-type specific hydroxymethylated regions can reveal cell-type specific enhancers[48].

1.2 Identifying tissue-specific histone marks

Given datasets of multiple histone modifications for a specific cell type, several tools, including ChromHMM[49,50], RFECS[51], and Segway[52], can define chromatin states across the cell's epigenome and/or define regulatory elements such as enhancers. While the above tools are designed for a single sample, tools like hiHMM[53] and TreeHMM[54] can define chromatin states in multiple cell types or multiple species simultaneously. But these tools cannot be readily applied to detect tissue or cell type-specific enhancers. Several efforts have been devoted to define tissue or cell type-specific enhancers. For example, the FANTOM5 Consortium identified active enhancers for a large number of human tissues and cell types by using bidirectional capped RNA data[55]. They called differentially expressed enhancers across all tissue and cell types, using Kruskal-Wallis rank sum tests. To define tissue differentially expressed enhancers, for example, for the brain, they further performed pair-wise, post-hoc tests, and required the enhancers to be differentially expressed between brain tissues and at least one non-brain tissue. A limitation of this approach is that such differentially expressed enhancers are often expressed in multiple tissue and cell types and are not specific to a single tissue or cell type. Furthermore, since the enhancers are marked by active transcription, poised enhancers are likely to be missed. Indeed, the active enhancers identified by FANTOM5 had 231 fold more bidirectional capped RNA reads than polycomb-repressed enhancers[55].

The Roadmap Epigenomics Project used a tool called HoneyBadger2 to define tissue or cell type-specific enhancers using k-means clustering. Regions that were clustered together share similar epigenetic profiles across a variety of tissue and cell types. A given cluster may have a pattern such that the enhancer signals are predominantly present in certain tissues, but not in other tissues. Such regions were defined as tissue-specific enhancers. However, this approach is based on unsupervised learning, and as such, clusters are not directly assigned to a specific tissue. Other groups characterized the cell-type specificity of enhancers in human and mouse using clustering methods[56–59], but did not provide tools to define cell-type specificity. Tools like MultiGPS[60] and dPCA[61] were designed to compare Chip-seq data between two conditions but not readily adaptable to compare enhancers or histone modifications between groups of tissue and cell types. Another tool, ChromDiff[62] compared chromatin states across different group of samples. For each given region, ChromDiff calculated the percent coverage for each chromatin state in each sample and corrected them based on sample metadata. Then it tested for difference of corrected values between two groups of samples for each chromatin state using statistical test such as Mann–Whitney–Wilcoxon test and identified significant regions with specific chromatin states. The tool can be applied to identify tissue or cell type-specific enhancers if ChromHMM models are defined, but can be difficult to use by experimental biologists due to the lack of a user-friendly interface.

1.3 Mapping DNA methylation and hydroxymethylation

Many genome-wide profiling methods were developed to map 5mC[63]. Technologies for mapping 5mC include bisulfite conversion of unmethylated cytosine to uracil (whole genome bisulfite sequencing (WGBS)), enrichment of methylated DNA using methyl-cytosine-specific antibodies (methylated DNA immunoprecipitation sequencing (MeDIP-seq)), and enrichment of unmethylated regions using methylation-sensitive restriction enzymes (methylation sensitive restriction enzyme sequencing (MRE-seq)) , followed by next-generation sequencing[63]. The gold standard method WGBS can measure methylation genome-wide at single-base resolution but requires high coverage of the genome (at least 10x coverage for each cytosine) and is 10 times more expensive than enrichment or restriction enzyme sequencing methods[64]. MeDIP-seq enriches for methylated regions and gives methylation status of enriched regions but has low resolution (about 150bp)[65,66]. MRE-seq can cover 30% of the genome using multiple restriction enzymes and give unmethylation status at single-base resolution for CpG at cut sites[65].

Similarly, 5hmC profiling technologies advanced from immunoprecipitation/enrichment-based methods to whole genome single-base resolution. Because WGBS cannot distinguish 5hmC from 5mC, Yu et. al. developed a method called TET-assisted bisulfite sequencing (TAB-seq), where 5hmCs are first protected by glucosylation and then 5mC is completely oxidized to 5caC with TET enzyme[46]. The following bisulfite treatment can reveal which CpGs are protected and infer hydroxymethylation levels. TAB-seq can measure genome-wide 5hmC at single-base resolution but requires very high coverage to confidently call 5hmC at all cytosines. For example, for 5% 5hmC, based on binomial test with a probability of 2.22% for 5mC non-conversion rate, a coverage of 120 is required to call 5hmC at 95% confidence level. The study from Yu, et al. could only confidently call 20% or higher 5hmC at an average coverage of 27. Often in TAB-seq experiments, both WGBS and TAB-seq libraries are deeply sequenced to parse out 5mC and 5hmC levels in a single sample. Achieving high confidence, single base resolution of 5hmC can be a heavy financial strain for large experimental designs due to the necessary sequencing depth. Therefore, many adopted the cheaper alternative of utilizing antibody-based enrichment method, such as hydroxymethylated DNA immunoprecipitation sequencing (hMeDIP-seq), which can reveal hydroxymethylated regions with limited sensitivity[45]. hMeDIP-seq can also provide relative hydroxymethylation over controls, but at the cost of low resolution. Similar to antibody-based enrichment method such as hMeDIP-seq, hmC-Seal chemically tags hydroxymethylated cytosine and enriches hydroxymethylated regions by pulling down tagged 5hmC[47,48]. hmC-Seal can pull down regions with extremely low 5hmc content and thus have higher sensitivity than hMeDIP-seq.

Because of the high cost of single-base-resolution profiling methods for 5hmC and 5mC, several computational methods were developed to estimate 5hmC and 5mC at single-base resolution. Xiao et.al. developed a random forest regression-based method MeSiC to estimate single-CpG 5mC from MeDIP-seq data[67]. Stevens et. al. took advantage of the complementary properties of MeDIP-seq and MRE-seq and developed a conditional random field-based algorithm methylCRF to effectively predict single-CpG 5mC from MeDIP-seq and MRE-seq data[68]. However, the two aforementioned algorithms cannot predict 5hmC levels. Pavlovic et al. developed a SVM/random forest-based method DIRECTION to predict single-CpG 5mC or 5hmC from histone modification and transcription factor ChIP-seq data[69]. This method can only predict binary values, either high or low 5mC/5hmC, but not the absolute quantitative level.

Chapter 2: EpiCompare: An online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features

Yu He, and Ting Wang

From:

EpiCompare: An online tool to define and explore genomic regions with tissue or cell type-specific epigenomic features. Bioinformatics 33, 3268–3275.

Publication at Bioinformatics, 2017

2.1 Abstract

The Human Reference Epigenome Map, generated by the Roadmap Epigenomics Consortium, contains thousands of genome-wide epigenomic datasets that describe epigenomes of a variety of different human tissue and cell types. This map has allowed investigators to obtain a much deeper and more comprehensive view of our regulatory genome, for example defining regulatory elements including all promoters and enhancers for a given tissue or cell type. An outstanding task is to combine and compare different epigenomes in order to identify regions with epigenomic features specific to certain types of tissues or cells, for example, lineage-specific regulatory elements. Currently available tools do not directly address this question. This need motivated us to develop a tool that allows investigators to easily identify regions with epigenetic features unique to specific epigenomes that they choose, making detection of common regulatory elements and/or cell type-specific regulatory elements an interactive and dynamic experience. An online tool EpiCompare was developed to assist investigators in exploring the specificity of epigenomic features across selected tissue and cell types. Investigators can design their test by choosing different combinations of epigenomes, and choosing different classification algorithms provided by our tool. EpiCompare will then identify regions with specified epigenomic features, and provide a quality assessment of the predictions. Investigators can interact with EpiCompare by investigating Roadmap Epigenomics data, or uploading their own data for comparison. We demonstrate that by using specific combinations of epigenomes we can detect developmental lineage-specific enhancers. Finally, prediction results can be readily visualized and further explored in the WashU Epigenome Browser.

2.2 Introduction

The Roadmap Epigenomics Consortium generated a reference catalogue of human epigenomes across a variety of tissue and cell types[70]. Using this resource, investigators can compare the epigenomes of different tissue and cell types and identify regulatory elements such as enhancers, promoters, and regions occupied by epigenetic features that are unique to a specific tissue or cell type, as well as those that are shared by multiple tissue and cell types.

One common application utilizing the Human Reference Epigenome is the identification of tissue or cell type-specific enhancers. Enhancers are cis-regulatory elements playing essential roles in regulating the spatial and temporal pattern of gene expression[71]. Many enhancers function in a tissue or cell type-specific manner[56–58]. Disruption of enhancer functions can often lead to diseases[72]. Many studies revealed that enhancers significantly overlap with disease-causal variants and such variants are often enriched in enhancers specific to cell types that are implicated in the specific diseases[56,70,73–77]. Hence, a comprehensive list of tissue or cell type-specific enhancers could have significant clinical impact.

The identification of tissue-specific histone marks including H3K27ac and H3K4me1 can help identify tissue or cell type-specific enhancers. Enhancers are epigenetically defined by the presence of H3K4me1 and the absence of H3K4me3[78]. H3K27me3 is a repression histone mark that is associated with polycomb complex[79]. The combination of H3K4me1 and H3K27me3 marks poised enhancers, which silence developmental genes in embryonic stem cells (ESCs) and keep them poised for activation in differentiating cells[80]. H3K27ac is a mark of active enhancers and promoters and distinguishes active enhancers from poised enhancers. Combination of H3K4me1 and H3K27ac modifications is used to identify active enhancers[81]. Therefore, combination of different histone marks can be used to predict tissue or cell type-specific poised/active enhancers.

Given datasets of multiple histone modifications for a specific cell type, several tools, including ChromHMM[49,50], RFECS[51], and Segway[52], can define chromatin states across the cell's epigenome and/or define regulatory elements such as enhancers. While the above tools are designed for a single sample, tools like hiHMM[53] and TreeHMM[54] can define chromatin states in multiple cell types or multiple species simultaneously. But these tools cannot be readily applied to detect tissue or cell type-specific enhancers. Several efforts have been devoted to define tissue or cell type-specific enhancers. For example, the FANTOM5 Consortium identified active enhancers for a large number of human tissues and cell types by using bidirectional capped RNA data[55]. They called differentially expressed enhancers across all tissue and cell types, using Kruskal-Wallis rank sum tests. To define tissue differentially expressed enhancers, for example, for the brain, they further performed pair-wise, post-hoc tests, and required the enhancers to be differentially expressed between brain tissues and at least one non-brain tissue. A limitation of this approach is that such differentially expressed enhancers are often expressed in multiple tissue and cell types and are not specific to a single tissue or cell type. Furthermore, since the enhancers are marked by active transcription, poised enhancers are likely to be missed. Indeed, the active enhancers identified by FANTOM5 had 231 fold more bidirectional capped RNA reads than polycomb-repressed enhancers[55].

The Roadmap Epigenomics Project used a tool called HoneyBadger2 to define tissue or cell type-specific enhancers using k-means clustering. Regions that were clustered together share similar epigenetic profiles across a variety of tissue and cell types. A given cluster may have a pattern such that the enhancer signals are predominantly present in certain tissues, but not in other tissues. Such regions were defined as tissue-specific enhancers. However, this approach is based on unsupervised learning, and as such, clusters are not directly assigned to a specific tissue. Other groups characterized the cell-type specificity of enhancers in human and mouse using clustering methods[56–59], but did not provide tools to define cell-type specificity. Tools like MultiGPS[60] and dPCA[61] were designed to compare Chip-seq data between two conditions but not readily adaptable to compare enhancers or histone modifications between groups of tissue and cell types. Another tool, ChromDiff[62] compared chromatin states across different group of samples. For each given region, ChromDiff calculated the percent coverage for each chromatin state in each sample and corrected them based on sample metadata. Then it tested for difference of corrected values between two groups of samples for each chromatin state using statistical test such as Mann–Whitney–Wilcoxon test and identified significant regions with specific chromatin states. The tool can be applied to identify tissue or cell type-specific enhancers if ChromHMM models are defined, but can be difficult to use by experimental biologists due to the lack of a user-friendly interface.

To address these needs, we have developed an online tool EpiCompare to help investigators to analyze the Roadmap Epigenomics data. Investigators can easily identify regions with epigenomic features specific to combinations of tissue or cell types. Several classification methods are provided, including the clustering method used by the Roadmap Epigenomics Project[70]. Investigators can compare enhancers, promoters, and specific histone marks using any combination of tissue and cell types, using Roadmap data and/or their own data. Investigators can test a variety of hypotheses by designing specific combinations of epigenome comparisons, and EpiCompare provides a quality assessment of the predictions. The predicted regions can be readily visualized and further explored within the WashU Epigenome Browser. EpiCompare makes Roadmap reference epigenomes more easily usable by experimental biologists in order to enhance their research.

2.3 Results

2.3.1 Performance comparison

To identify regions with epigenomic features specific to combinations of tissue or cell types, we applied three different methods: frequency cutoff, Fisher's exact test, and k-means clustering, as described in Methods. The most important parameters for all the methods are choices of foreground samples and background samples (see Methods). The main assumption we make is that the epigenomic features we focus on are enriched in foreground samples but depleted in background samples. Identified regions were tested using the following validation methods: GREAT analysis, enrichment for DNase I hypersensitive sites (DHS) and H3K27ac peaks, and the tissue enrichment index, contribution measure (CTM) (see Supplementary Note 1). CTM measures how much a sample or a group of samples contributes to the total amount of signal (e.g., read density for H3K27ac) combined by all samples in a region[82]. To further evaluate the performance directly, we randomly picked 20 identified regions and visualized them in WashU Epigenome Brower with chromatin states and histone modification tracks. We used adult brain tissues as foreground samples and evaluated the efficacy of the three methods in identifying adult brain-specific enhancers using enhancers defined by 15-state ChromHMM model. Seven adult brain samples were available from the Roadmap Epigenomics Project. We compared them to 91 other samples with available H3K27ac data. Since the clustering method does not provide ranks, we obtained a list of adult brain-specific enhancers using the clustering method with default settings. We then picked an equal number of regions in ascending order of ranks using the frequency cutoff and Fisher's exact test methods.

First, we examined the overlap of enhancers found by three methods (**Supplementary Fig. S2**). Out of 188,076 identified adult brain-specific enhancers (i.e., 200bp windows), 148,170 overlapped between the frequency cutoff and Fisher's exact test; 133,370 overlapped between k-means clustering and Fisher's exact test; and 144,182 overlapped between frequency cutoff and k-means clustering. 123,746 were shared across all three methods.

Next, we tested our predicted brain-specific enhancers using the three validation methods. Using the GREAT[83], we found that adult brain-specific enhancers identified by each of three methods were strongly associated with brain functions such as myelination, regulation of action potential and regulation of synaptic plasticity (**Figure. 1(a)** and Supplementary **Fig. S3**). The brain-specific enhancers predicted by all three methods also had much higher enrichment for H3K27ac peaks in brain tissues compared to other tissues (**Figure. 1(b)** and **Supplementary Fig. S4**). Overall, the enrichment in brain tissues was higher for the frequency cutoff and Fisher's exact test methods than for the clustering method. The brain-specific enhancers predicted by all three methods also had much higher CTM index in brain tissues than in other tissues for H3K27ac-based CTM distribution (**Figure. 1(c)** and **Supplementary Fig. S5**), underscoring the brain specificity of the enhancer histone modification in the identified regions. The brain tissue CTM distributions for regions identified by the three methods almost superimposed each other (**Supplementary Fig. S5**). A visualization of randomly picked 20 brain-specific enhancers identified from Fisher's exact test showed most regions had much stronger H3K4me1/H3K27ac peaks in the foreground samples than the background samples (**Supplementary Fig. S6**). In summary, the validation results confirmed that our methods can effectively identify tissue-specific enhancers. Similarly, the same methods can be applied to identify other epigenomic modifications that are tissue or cell-type specific.

Since FANTOM5 defined active enhancers for a variety of tissue and cell types by their differential expression patterns, we compared brain-specific enhancers identified by Fisher's exact test on enhancers defined by 15-ChromHMM model and the FANTOM5. First, we examined the overlap between these two methods. The FANTOM5 enhancers were not binned on 200bp windows, so we mapped them onto 200bp windows. 89 of 208,804 regions by ChromHMM-based method overlapped with 1,578 binned FANTOM5 brain enhancers (hypergeometric test, p=10-26). The overlap was small because only 11% of H3K4me1/H3K27ac loci overlapped the FANTOM5 enhancers[55], and enhancers defined by ChromHMM included active enhancers (characterized by H3K4me1/H3K27ac loci), poised enhancers (characterized by H3K4me1/H3K27me3 loci), and other types of enhancers (single H3K4me1 mark or single H3K27ac mark). Although the overlap with the FANTOM5 brain enhancers was small, it was highly significant. In contrast, 35 regions by ChromHMM-based method overlapped with 3,409 binned FANTOM5 blood enhancers (hypergeometric test, p=0.98), suggesting the overlap was specific to brain. Second, we plotted enrichment of H3K27ac for the shared regions, as well as regions unique to each method (**Supplementary Fig. S7**). Finally, we randomly picked 20 regions that were unique to each method, and visualized them on the WashU Epigenome Browser in gene set view (**Supplementary Fig. S8** and **S9**). Interestingly, we found that many FANTOM5-defined brain-specific enhancers are defined as promoters by using Roadmap Epigenomics data, with clear and strong promoter histone mark support (i.e., H3K4me3). Moreover, these regions also have high H3K27ac in the background samples.

Using similar analysis as above, we compared identifying brain-specific enhancers using Fisher's exact test and ChromDiff (See Supplemental Note 9). We found enhancers identified from Fisher's exact test and ChromDiff largely overlapped (80%). Enhancers that were unique to Fisher's exact test had much stronger enrichment of H3K27ac in brain samples than ChromDiff but also had higher enrichment in the background samples. Therefore anecdotally EpiCompare seems to have better sensitivity, while ChromDiff seems to exhibit better specificity, at a comparable statistical cutoff. ChromDiff is a command line only program, while EpiCompare provides a much more user-friendly interface and includes access to WashU Epigenome Browser, allowing biologists to better explore their result.

The k-means clustering method in our tool is similar to the clustering method used in HoneyBadger2 tool with the exception that enhancers defined by the 15-state ChromHMM model in HoneyBadger2 were further filtered by DHS before used for clustering. To demonstrate that our clustering method is comparable to HoneyBadger2, we compared adult brain-specific enhancers identified by the two approaches. We used 250 clusters as a close approximation of 246 clusters in HoneyBadger2 tool. We identified 158,110 regions with our approach and 86,019 regions with HoneyBadger2. For the comparison, we randomly picked 86,019 regions from the total regions identified by our approach. By comparing the enrichment of H3K27ac peaks in the foreground samples and background samples between our clustering method and HoneyBadger2, we found that both methods had similar enrichment in the foreground samples (t-test, p=0.87) and also in the background samples (t-test, p=0.98) (**Supplementary Fig. S10(a)**). When we examined the CTM distribution of H3K27ac, we found that the brain tissue CTM distributions for regions identified by the two methods almost superimposed each other (**Supplementary Fig. S10(b)**). Thus our clustering method is comparable to the clustering method in HoneyBadger2 tool.

After demonstrating that our methods can identify tissue-specific enhancers, we determined the impact of sample size on performance: i.e., the impact of the number of foreground samples and the number of background samples (see Supplementary Note 2). First, to examine how the number of foreground samples affects the performance, we predicted adult brain-specific enhancers by using different number of foreground samples while fixing background samples. To assess performance, we computed the average enrichment of H3K27ac peaks in the seven adult brain samples and also in selected background samples because we expect that tissue-specific enhancers should have higher enrichment in the foreground samples and lower enrichment in the background samples. We found that with increasing foreground samples, the performance of all three methods increased. This is illustrated by increasing H3K27ac enrichment in the foreground samples, and relatively stable depletion in the background sample (**Figure. 2(a)**).

To examine how the number of background samples affects the performance, we predicted adult brain-specific enhancers by using different number of background samples while fixing foreground samples. The enrichment of H3K27ac in the foreground samples seemed to be quite stable across a range of numbers of background samples used (**Figure. 2(b)**). However, depletion of H3K27ac in background samples seemed to be quite sensitive to the number of background samples used A larger number of background samples did improve the specificity effectively, underscoring the importance of having a comprehensive collection of epigenomes, such as those made available by the Roadmap Epigenomics project.

Finally, we demonstrate that our simple but versatile framework allows investigators to design any combination of epigenome comparison to identify specific epigenomic features associated with specific biological entities. For example, by combining samples that share the same developmental origin, one might be able to identify specific regulatory mechanisms for this developmental lineage. This is particularly useful when samples representing cells in early development are difficult to obtain. Here we set out to define endoderm-specific enhancers by comparing nine adult tissues derived from the endoderm to other background tissues (see Supplementary Note 3). The enhancers were defined using 18-state ChromHMM model. We identified 13,728 regions using frequency cutoff method, 46,859 regions using Fisher's exact test method, and 29,386 regions using k-means clustering method with 140 clusters. We picked top 13,728 from Fisher's exact test for the following analysis. The predicted regions exhibited much stronger enrichment of DHS in endoderm-derived tissues than in other tissues (**Figure. 3(a)** and **Supplementary Fig. S11**). Moreover, when subjected to analysis by the GREAT tool, these regions were strongly associated with biological processes related to epithelial cell functions (**Figure. 3(b)**), a well-known derivative function common for endoderm-derived tissues[84]. A visualization of randomly picked 20 endoderm-specific enhancers identified from Fisher's exact test showed most regions had much stronger H3K4me1/H3K27ac peaks in the foreground samples than the background samples (**Supplementary Fig. S12**). To further explore the functions of these endoderm-specific enhancers, we identified potential regulatory transcription factors (TFs) interacting with these regions by HOMER[85]. The top enriched TFs are all important for endoderm specification, including FoxA family TFs (FoxA1, FoxA2), GATA family TFs (Gata4), HNF1, HNF4a and others (**Figure. 3(c)**). FoxA family and GATA family TFs are key players in the transcriptional regulatory network of the endoderm[84]. FoxA1 and FoxA2 are pioneer TFs that remodel chromatin environment and facilitate recruitment of other TFs[86]. FoxA1 and FoxA2 are homologous and required for the development of endoderm tissues such as liver, lung, intestine and pancreas[87–90]. Like FoxA1 and FoxA2, HNF1 and HNF4a play key regulatory roles in liver, pancreas, and intestine development[91–93]. Moreover, the foregut markers PDX1 and the hindgut marker CDX2 were also highly enriched (p= 10-5 for PDX1 and CDX2 motifs)[94]. To further support the function of the top enriched TFs in endoderm tissues, many of them were highly expressed in endoderm tissues comparing to non-endoderm tissues, ESCs and ESC-derived multipotent endoderm cells[70] (**Supplementary Fig. S13**).

Using top enriched TFs that were also highly expressed in adult endoderm tissues compared to non-endoderm adult tissues, we identified 4 upstream TF candidates - FoxA1, FoxA2, HNF1b, HNF4a, and were able to build a transcriptional regulatory network for them and shared target genes by linking enhancers with TF binding sites to nearest genes (**Figure. 3(d)** and **Supplementary Table S1**) using previously described methods[95]. The reconstructed network recapitulated many important gene regulation relationships in endoderm development and differentiation. For example, the FoxA family TFs cooperate with HNF1b and HNF4a to regulate intestinal epithelial cell function[93]. FoxA2, HNF1b, and HNF4a were shown to bind to a large number of target regions in intestinal epithelial cell line[93]. The 72 shared target genes for the 4 TFs were enriched for signaling pathways required for cell proliferation and differentiation including WNT, BMP, VEGF and Hippo signaling (**Supplementary Table S2**)[96]. The median expression level of these genes was significantly higher in endoderm tissues than that in non-endoderm tissues (t-test, p=5e-5) (**Supplementary Fig. S14**). To further confirm that the network was activated in endoderm tissues, we examined the profile of epigenetic marks (DNase I, H3K27ac and DNA methylation) on all enhancers in this network across different tissues, including adult endoderm tissues, fetal endoderm tissues, endoderm cells, non-endoderm tissues and ESCs. These enhancers showed strong expression of DNase I and H3K27ac mark and low DNA methylation only in adult endoderm tissues and fetal endoderm tissues (**Figure. 3(e)**). **Figure 3(f)** gave an example of merged endoderm-specific enhancers. The enhancers had strong DHS and H3K27ac peaks and low DNA methylation level in both adult and fetal endoderm tissues but not others. The evidence suggests that this regulatory cascade is active in fetal and adult endoderm tissues, but not in ESC-derived endoderm cells which presumably have not committed to a special endoderm cell type and also not in non-endoderm tissues.

2.3.2 Web server

The tool EpiCompare is freely available online. It was written in R using the Shiny framework and hosted by open source shiny server[97]. The home page includes a simple and intuitive user interface for the selection of foreground samples and background samples from a list of human tissue and cell types available from the Roadmap Epigenomics Consortium (**Supplementary Fig. S15**). Options for selecting different classification methods and parameters are also provided. It also provides the option of uploading user's data for analysis. The results page provides analysis results, including H3K27ac enrichment and tissue enrichment index using H3K27ac expression data. Results are presented as a table of identified regions, and can be downloaded for further analysis. Each region is linked to the WashU Epigenome Browser[98] where users can visualize, explore, and compare their epigenomic patterns in different tissue/cell types. The help page gives a tutorial on how to use EpiCompare.

2.4 Discussion and Conclusions

We have developed an online tool EpiCompare to help investigators to analyze the Roadmap Epigenomics data. The presented data showed that the tool can easily identify regulatory elements such as enhancers, promoters, and regions occupied by epigenetic features that are unique to a specific tissue or cell type, as well as those that are shared by multiple tissue and cell types. Our tool is designed specifically for biologists in such a way that no programming or data processing capacity is required to perform genome-wide analysis. We demonstrated that our tool could identify endoderm-specific enhancers and analysis on these enhancers revealed the regulatory network common to all endoderm tissues.

In identifying regions with epigenomic features specific to combinations of tissue or cell types, EpiCompare has several advantages over existing methodologies reported in the FANTOM5, Roadmap, and others. First, investigators can compare enhancers, promoters, and specific histone marks using any combination of tissue and cell types depending on their needs. This enables the identification of specific epigenomic features associated with specific biological entities, such as lineage-specific enhancers. Second, the tool is user-friendly so that an experimental biologist with little or no programming experience can easily use. Investigators can test a variety of hypotheses by designing specific combinations of epigenome comparisons using Roadmap data and/or their own data, and EpiCompare provides a quality assessment of the predictions. The predicted regions can be readily visualized and further explored using the WashU Epigenome Browser.

EpiCompare has some limitations. First, the regulatory elements used in this tool are defined based on the ChromHMM model. Although considered the state-of-the-art, ChromHMM model still has limited sensitivity and specificity, especially for identifying enhancers[76]. The performance of predicting tissue or cell type-specific enhancers is clearly dependent on the performance of ChromHMM. Second, EpiCompare is based on comparison of binary data including chromatin states and histone mark peaks. It could potentially miss regions with quantitatively different signal between samples. For example, it could not distinguish a weak enhancer from a strong enhancer if both had signals over the threshold. It could also not distinguish two quantitatively different weak enhancers which were below the threshold. These cases are false negatives for EpiCompare. The comparison of binary data can also lead to false positives if two samples had very similar signal at one region, with one above the threshold and the other below the threshold. Third, we implemented three very simple statistical models, and potentially could oversimplify the problem of identifying tissue or cell type-specific features. Frequency cutoff method uses simple cutoffs, and Fisher's exact test assumes the occurrence of features as hypergeometric distribution while k-means clustering method assumes certain number of clusters in the data and groups them based on similarity. All of them assume the independence of samples, but biological samples are clearly not independent from each other. The statistical models also do not consider the distribution of each feature along the genome of each sample. However, we are encouraged by the strong performance of these simple models, and anticipate that development of more sophisticated models will surely improve the accuracy of feature identification.

2.5 Methods

2.5.1 Datasets

The Roadmap Epigenomics Consortium uses the ChromHMM tool to generate chromatin states for different tissue and cell types. The type and number of chromatin states depends on the histone modification data provided. The 15-state ChromHMM model integrates histone modifications H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K36me3, while the 18-state ChromHMM model integrates the five marks in the 15-state model plus H3K27ac[70]. From the Roadmap Epigenomics Project, we obtained 15-state and 18-state ChromHMM models, and processed peak data (obtained from MACS[99]) for H3K27ac, H3K4me1, H3K4me3 and H3K27me3 marks for all tissue and cell types. Chromatin states are predicted for each 200 base pair (bp) window. The 15-state ChromHMM model defines enhancers as state numbers 6, 7, 12, corresponding to genic enhancers, enhancers, and bivalent enhancers, respectively. The 18-state ChromHMM model defines enhancers as state numbers 7, 8, 9, 10, 11, 15, corresponding to genic enhancer 1, genic enhancer 2, active enhancer 1, active enhancer 2, weak enhancer, and bivalent enhancer, respectively. Further, for all processed peak data, the coordinates are mapped to 200bp windows by requiring at least 50bp overlapping. Only peaks with q-value less than 0.01 are considered. Each feature above – the enhancer state or epigenomic modification peak – is converted into binary presence or absence of the feature in each 200bp window, denoted by 1 or 0. A table is generated for each feature by summarizing the presence or absence of the feature in all samples across windows where at least one sample has the feature.

2.5.2 Classification methods

EpiCompare contains three methods for identifying regions with epigenomic features specific to combinations of tissue or cell types (**Supplementary Fig. S1**). All methods require the definition of foreground samples and background samples by users. Foreground samples are the group of samples for which we identify specific regions. Background samples are the group of samples against which we compare foreground samples. The principle of all methods is, to define regions with features specific in foreground samples, the features should be enriched in the foreground samples but depleted in the background samples.

The first method implements a frequency cutoff. For each region (in this case each 200bp genomic window), the percentages of samples having the feature in the foreground samples and background samples are calculated. If the percentage of samples having the feature in the foreground samples is greater than or equal to the defined minimal foreground cutoff (default is 80%) and the percentage of samples having the feature in the background samples is less than or equal to the defined maximal background cutoff (default is 20%), then the region is defined as a positive region. These positive regions are further ranked by the difference between the percentage of samples having the feature in the foreground samples and background samples so users can prioritize top-ranked regions.

The second method implements Fisher's exact test. For each 200bp window, a contingency table composed of the number of samples with or without the feature in foreground samples and background samples is calculated. Fisher's exact test is used to examine whether the percentage of features in the foreground samples is significantly greater than in the background samples. The p-value is corrected by multiple hypothesis testing using the Benjamini-Hochberg procedure, and regions with q-value less than a cutoff (default is 0.01) are identified and ranked by their q-values. The statistical power of the test depends on the number of foreground samples and background samples and having more samples can provide more statistical power to identify more significant q-values (See Supplementary Note 6). Therefore, when the number of foreground samples or background samples is small, investigators can use q-value as a ranking measure and obtain the top candidates by setting a higher q-value threshold. We also evaluated the false positive rate of Fisher's exact test (See Supplementary Note 7).

The third method implements k-means clustering based on a Jaccard-index distance, similar to the clustering method used in HoneyBadger2[70]. First, k-means clustering is performed on regions in the binary data table for each feature. R package flexclust is used for clustering[100]. We determined the optimal cluster number by the elbow method and the silhouette method[101] (See Supplementary Note 8). The optimal cluster number for all features is close and around 140, so we provide the optimal cluster number for all features to be 140. In addition to the default number, we provide several other options (i.e., cluster number 90, 200 and 250) to give users flexibility. Next, the percentage of regions having the feature is calculated for each cluster and defined as a feature density table (number of clusters times number of samples). Finally, a cluster specific for a tissue/cell type should have higher feature density in that tissue/cell type than in the background samples. Specifically, to identify clusters specific for foreground samples, we select clusters satisfying the following two conditions: first, the median of feature densities of foreground samples in a cluster is greater than or equal to a threshold (default is 0.4); second, it should also be greater than or equal to the highest feature density in the background samples of that same cluster (this threshold can be set to any percentile of feature densities in the background samples).

2.6 Acknowledgements and Funding

The authors thank Dr. Bo Zhang and all members in Ting Wang's lab for suggestions on the manuscript and the website. This work was supported by the National Institutes of Health [R01HG007354, R01HG007175, R01ES024992, U01CA200060, U24ES026699, U01HG009391]; and American Cancer Society Research Scholar [RSG-14-049-01-DMC].

2.7 Author Contributions

Y.H developed the tool and did all analysis. Y.H and T.W designed the study and wrote the manuscript.

2.8 Figures & Tables

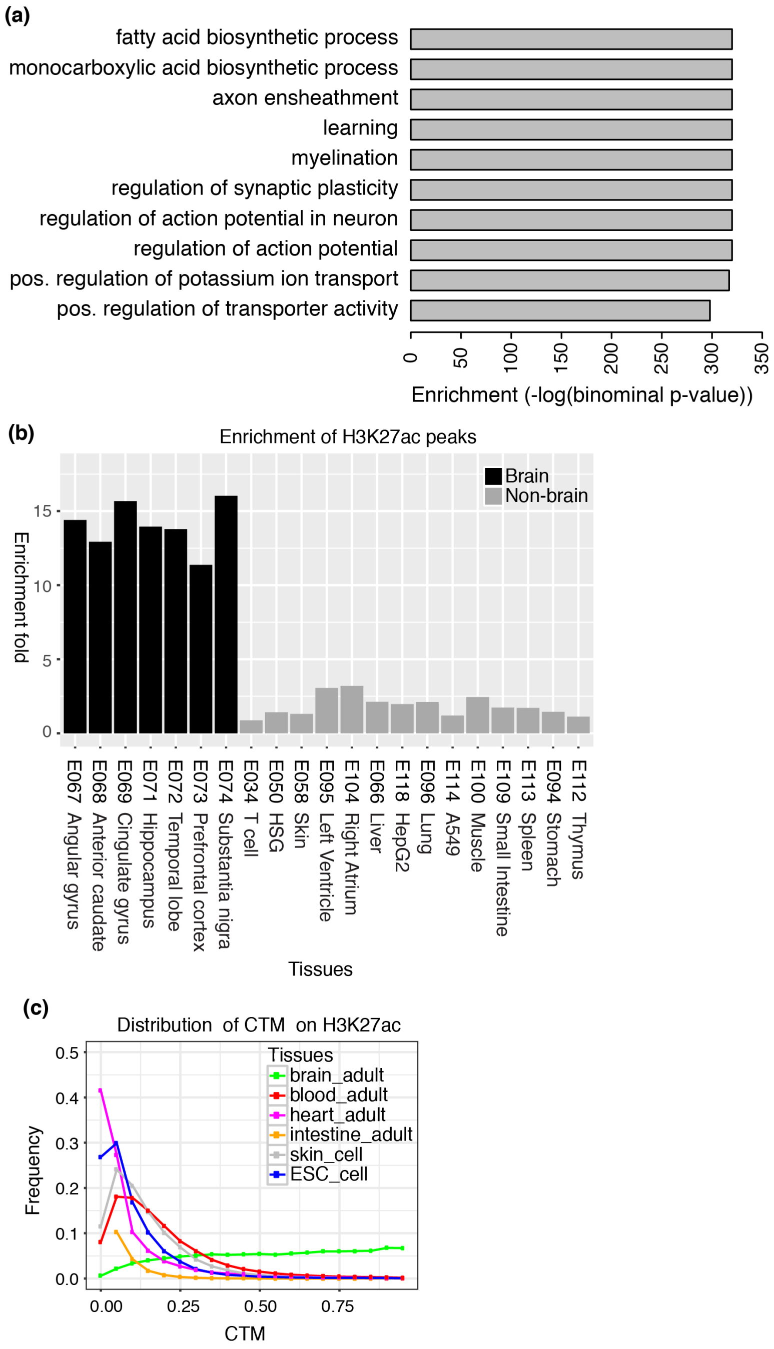


Figure 1. Validation of predicted brain-specific enhancers by Fisher's exact test method.

(**a**) Enriched GO terms and their binomial p-values based on GREAT. The top 10 GO terms are displayed here. (**b**) Enrichment of H3K27ac peaks in brain tissues and non-brain tissues for predicted adult brain-specific enhancers by Fisher's exact test. (**c**) The distribution of tissue enrichment index CTM based on H3K27ac expression data for predicted adult brain-specific enhancers by Fisher's exact test.

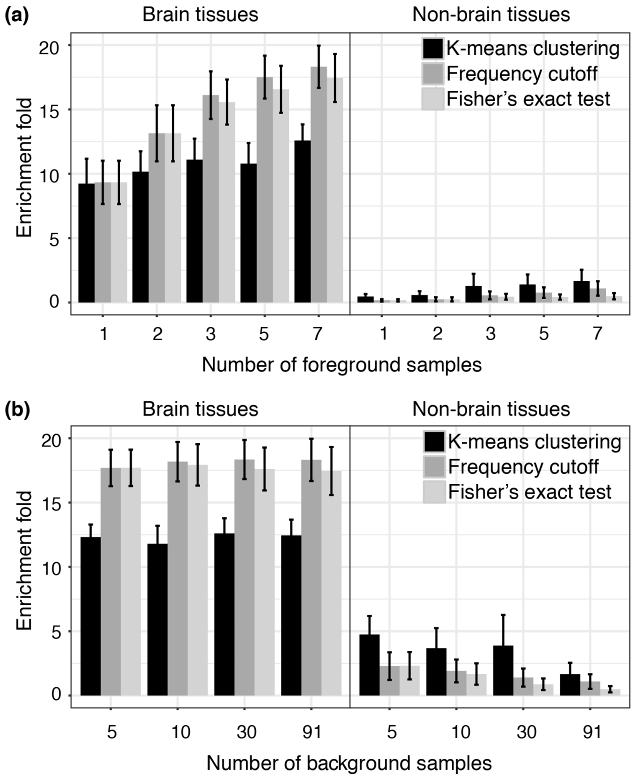


Figure 2. The effect of sample size on the performance of adult brain specific-enhancer predictions.

(**a**) How the number of foreground samples influences the performance with fixed background samples. (**b**) How the number of background samples influences the performance with fixed foreground samples.

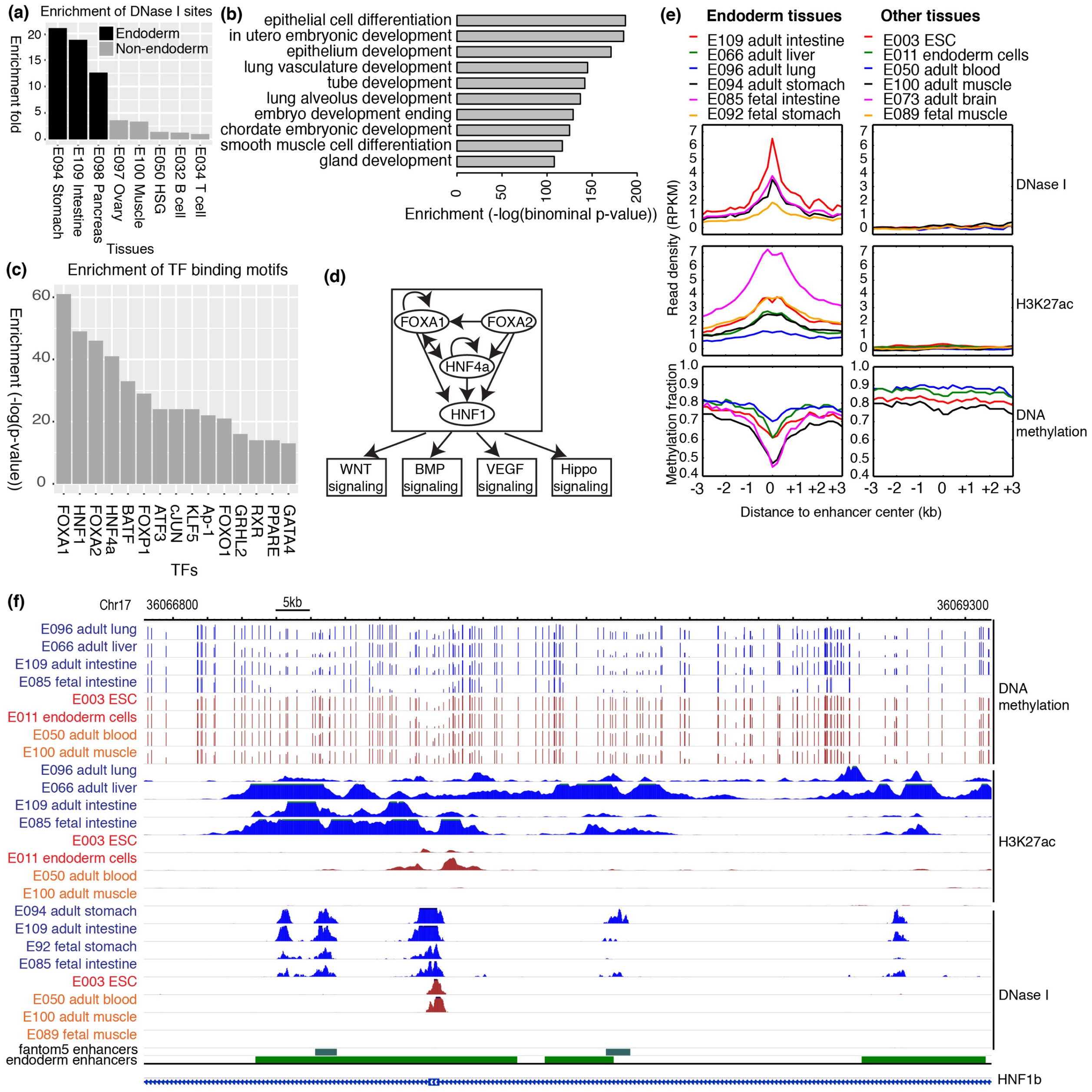
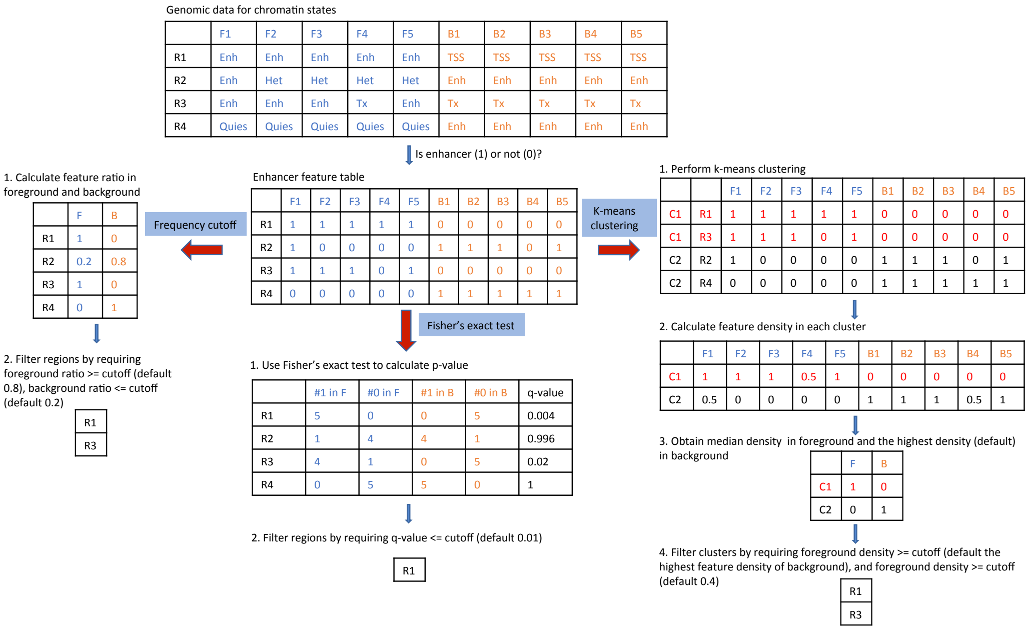


Figure 3. Identification of endoderm-specific enhancers by Fisher's exact test method.

(**a**) Enrichment of DHS for endoderm-specific enhancers identified by Fisher's exact test. (**b**) Enriched GO terms and their binomial p-values based on GREAT. Top 10 terms are displayed. (**c**) Enrichment of TF binding motifs in endoderm-specific enhancers by Fisher's exact test. Top 15 TFs are displayed. (**d**) The putative gene regulatory networks for endoderm tissues based on identified enhancers. (**e**) The expression profiles of epigenetic marks for enhancers in the network in endoderm tissues and non-endoderm tissues, ESCs and ESC-derived endoderm cells. (**f**) A browser example of merged endoderm-specific enhancers. Blue is endoderm tissues, brown is non-endoderm tissues, and red is ESCs and ESC-derived endoderm cells.



Genomic data for chromatin states

	F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
R1	Enh	Enh	Enh	Enh	Enh	TSS	TSS	TSS	TSS	TSS
R2	Enh	Het	Het	Het	Het	Enh	Enh	Enh	Enh	Enh
R3	Enh	Enh	Enh	Tx	Enh	Tx	Tx	Tx	Tx	Tx
R4	Quies	Quies	Quies	Quies	Quies	Enh	Enh	Enh	Enh	Enh

Is enhancer (1) or not (0)?

Enhancer feature table

	F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
R1	1	1	1	1	1	0	0	0	0	0
R2	1	0	0	0	0	1	1	1	0	1
R3	1	1	1	0	1	0	0	0	0	0
R4	0	0	0	0	0	1	1	1	1	1

1. Calculate feature ratio in foreground and background

	F	B
R1	1	0
R2	0.2	0.8
R3	1	0
R4	0	1

2. Filter regions by requiring foreground ratio >= cutoff (default 0.8), background ratio <= cutoff (default 0.2)

R1
R3

1. Use Fisher’s exact test to calculate p-value

	#1 in F	#0 in F	#1 in B	#0 in B	q-value
R1	5	0	0	5	0.004
R2	1	4	4	1	0.996
R3	4	1	0	5	0.02
R4	0	5	5	0	1

2. Filter regions by requiring q-value <= cutoff (default 0.01)

R1

1. Perform k-means clustering

		F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
C1	R1	1	1	1	1	1	0	0	0	0	0
C1	R3	1	1	1	0	1	0	0	0	0	0
C2	R2	1	0	0	0	0	1	1	1	0	1
C2	R4	0	0	0	0	0	1	1	1	1	1

2. Calculate feature density in each cluster

	F1	F2	F3	F4	F5	B1	B2	B3	B4	B5
C1	1	1	1	0.5	1	0	0	0	0	0
C2	0.5	0	0	0	0	1	1	1	0.5	1

3. Obtain median density in foreground and the highest density (default) in background

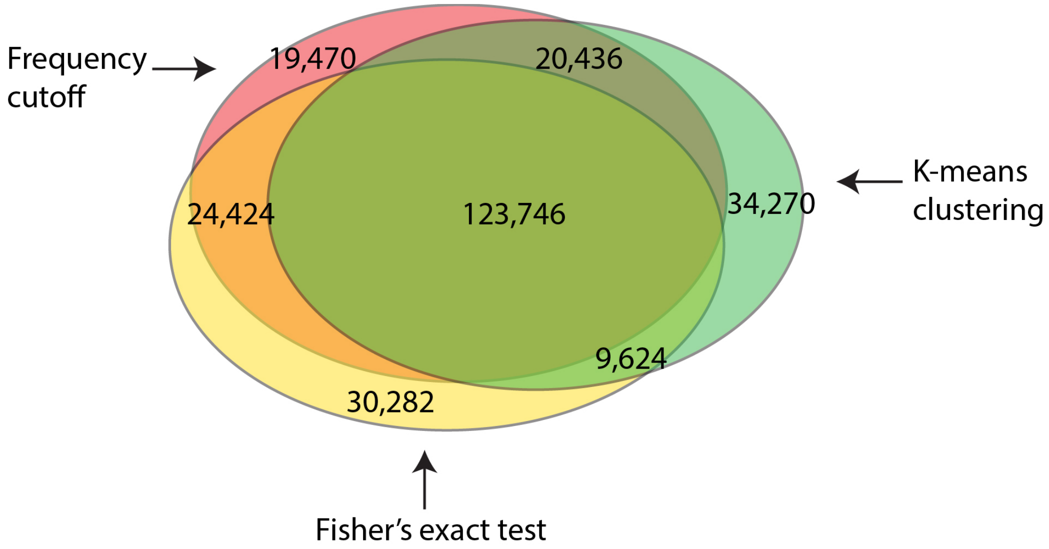
	F	B
C1	1	0
C2	0	1

4. Filter clusters by requiring foreground density >= cutoff (default the highest feature density of background), and foreground density >= cutoff (default 0.4)

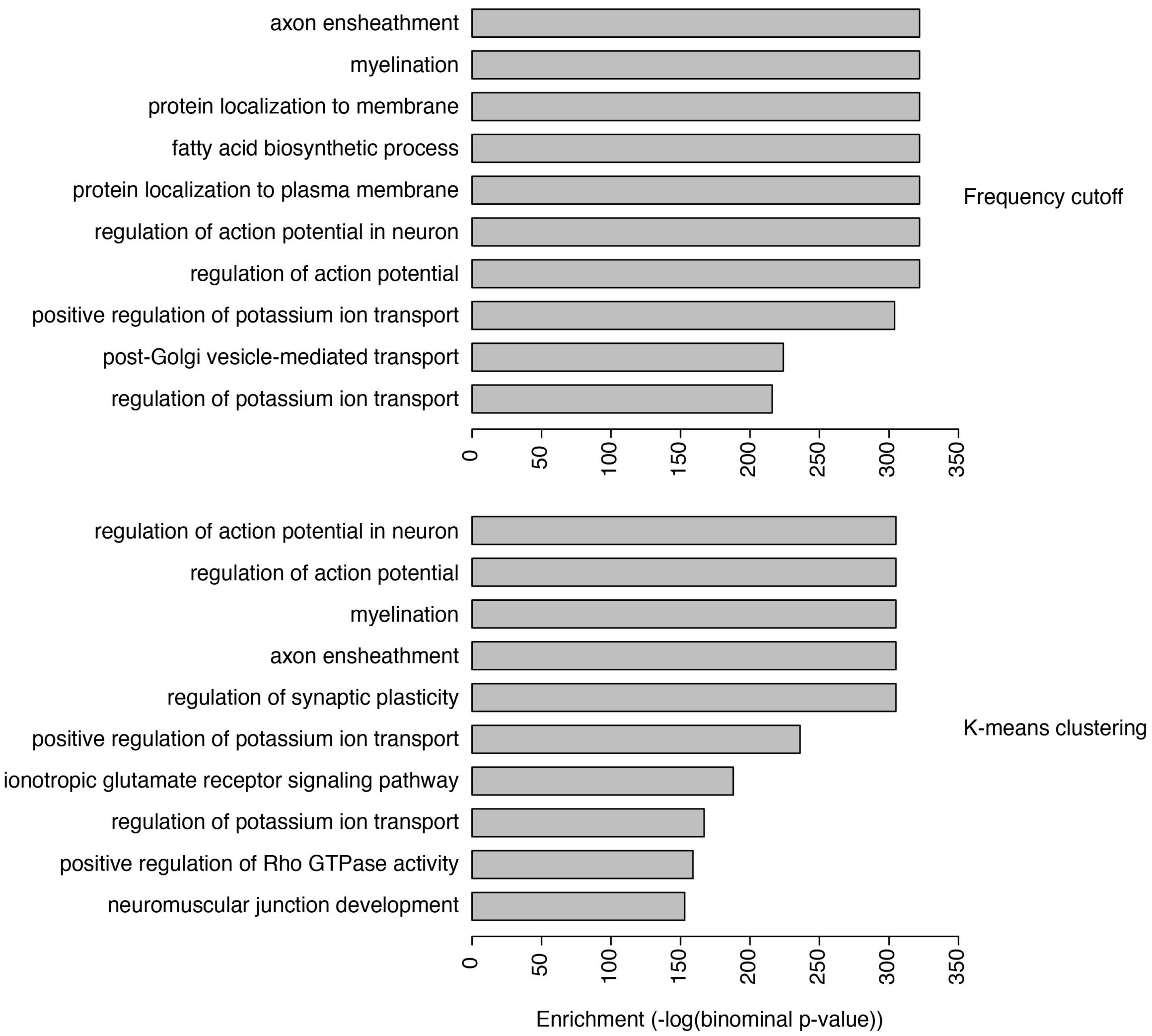
R1
R3

Supplementary Figure 1. Visualization of three methods in EpiCompare with a simple example.

F represents foreground samples, B represents background samples, R represents regions, C represents clusters.

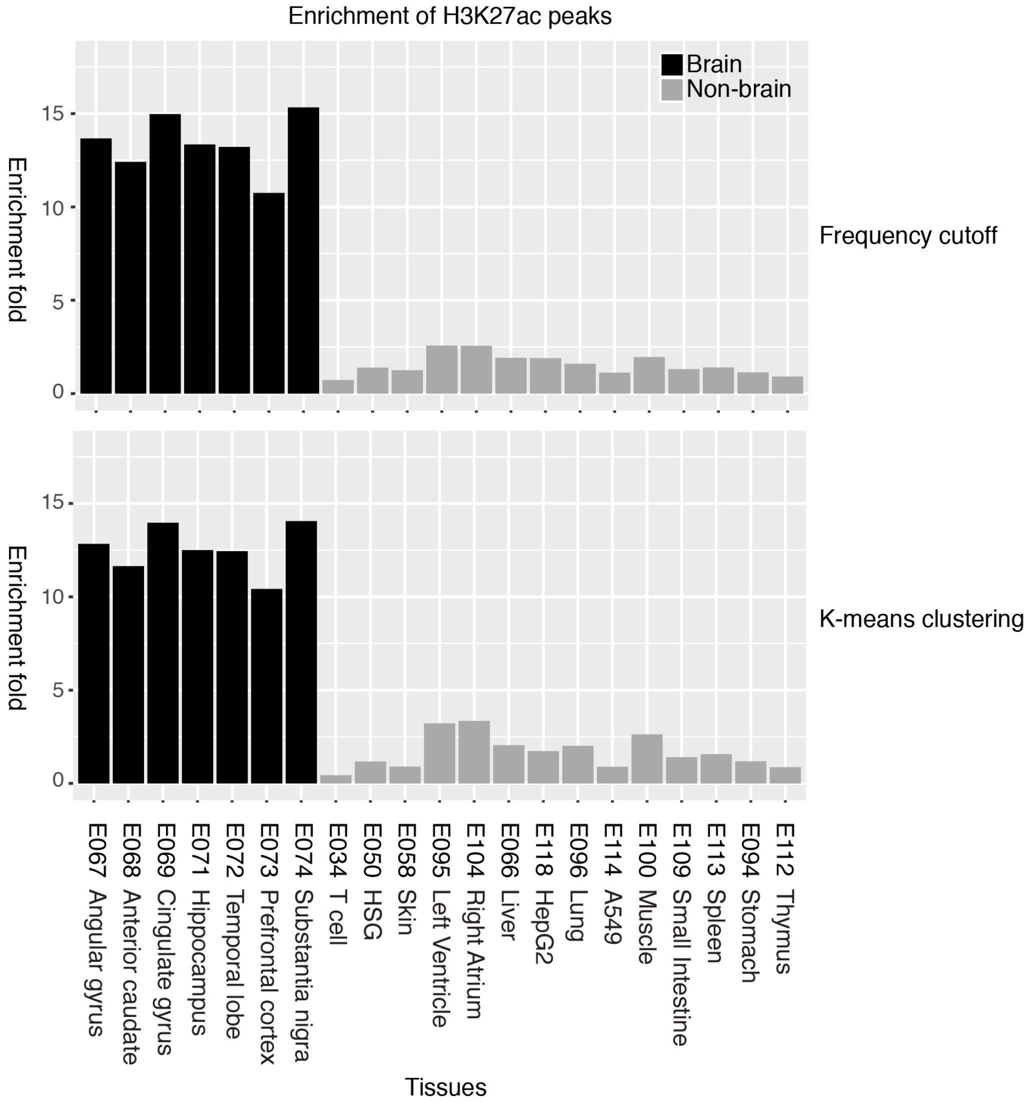


Supplementary Figure 2. The overlap of adult brain-specific enhancers identified by three methods.

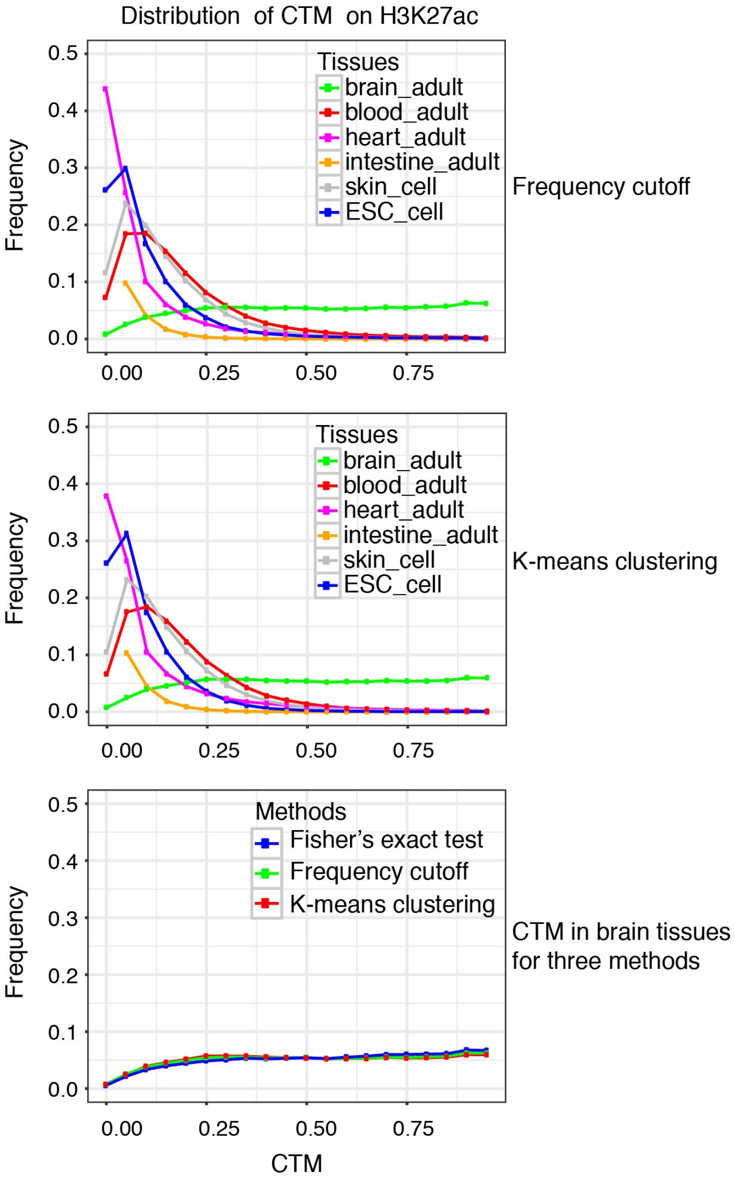


Supplementary Figure 3. Enriched GO terms and their binomial p-values from analyzing predicted adult brain-specific enhancers by frequency cutoff and k-means clustering method based on GREAT.

The top 10 GO terms are displayed here.

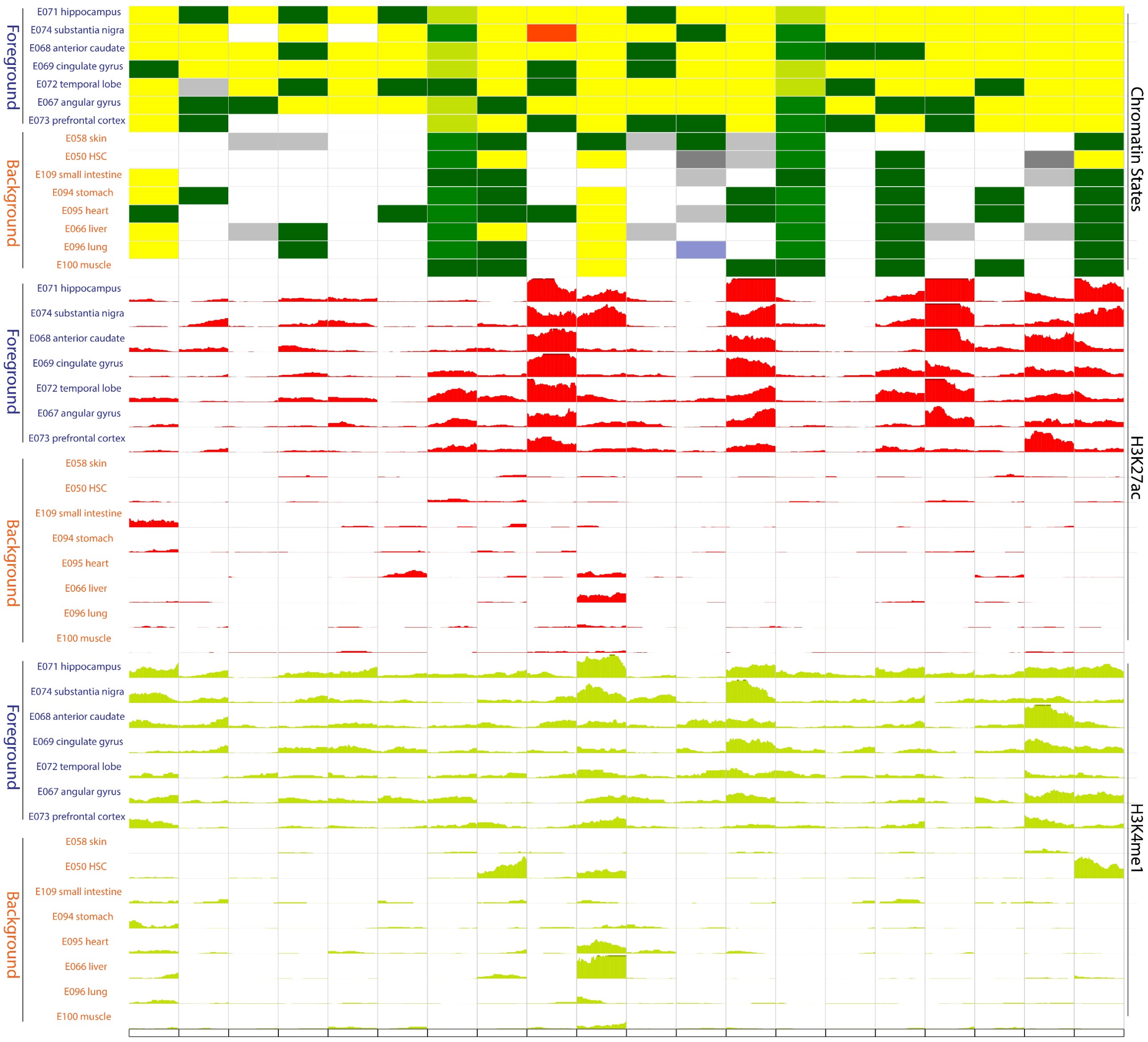


Supplementary Figure 4. Enrichment of H3K27ac peaks in brain and non-brain tissues for adult brain-specific enhancers identified by frequency cutoff and k-means clustering method.



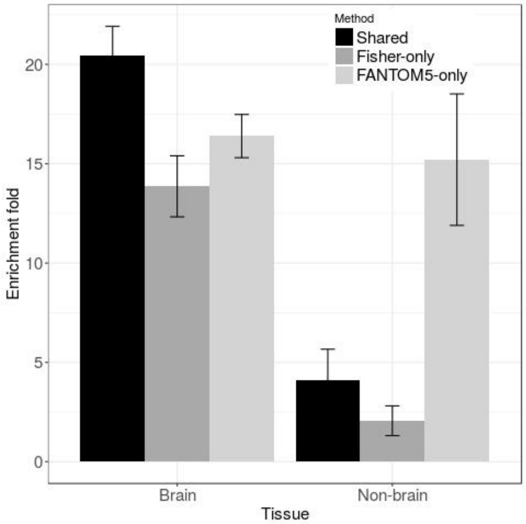
Supplementary Figure 5. The distribution of tissue enrichment index CTM based on H3K27ac RPKM data for adult brain-specific enhancers identified by frequency cutoff and k-means clustering method.

For comparison, the figure in the last row merges CTM index distribution in brain tissues for Fisher's exact test, frequency cutoff and k-means clustering method.

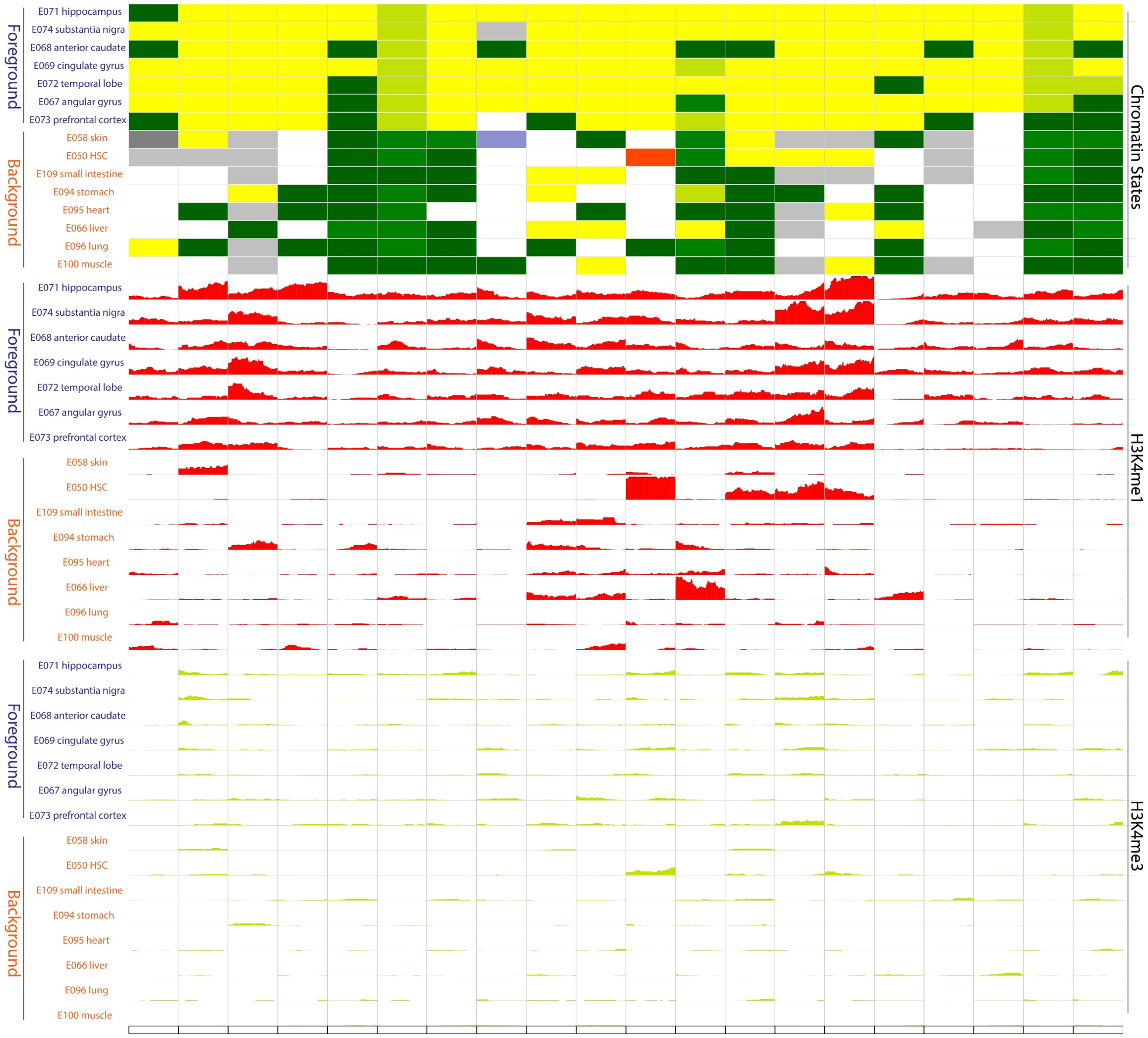


Supplementary Figure 6. Visualization of randomly picked 20 brain-specific enhancers identified from Fisher's exact test

The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 15 state, H3K27ac signal track (-log10(p-value)), H3K4me1 signal track (-log10(p-value)). Yellow color in ChromHMM state track represents enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Foreground samples are labeled as blue and background samples are labeled as brown.

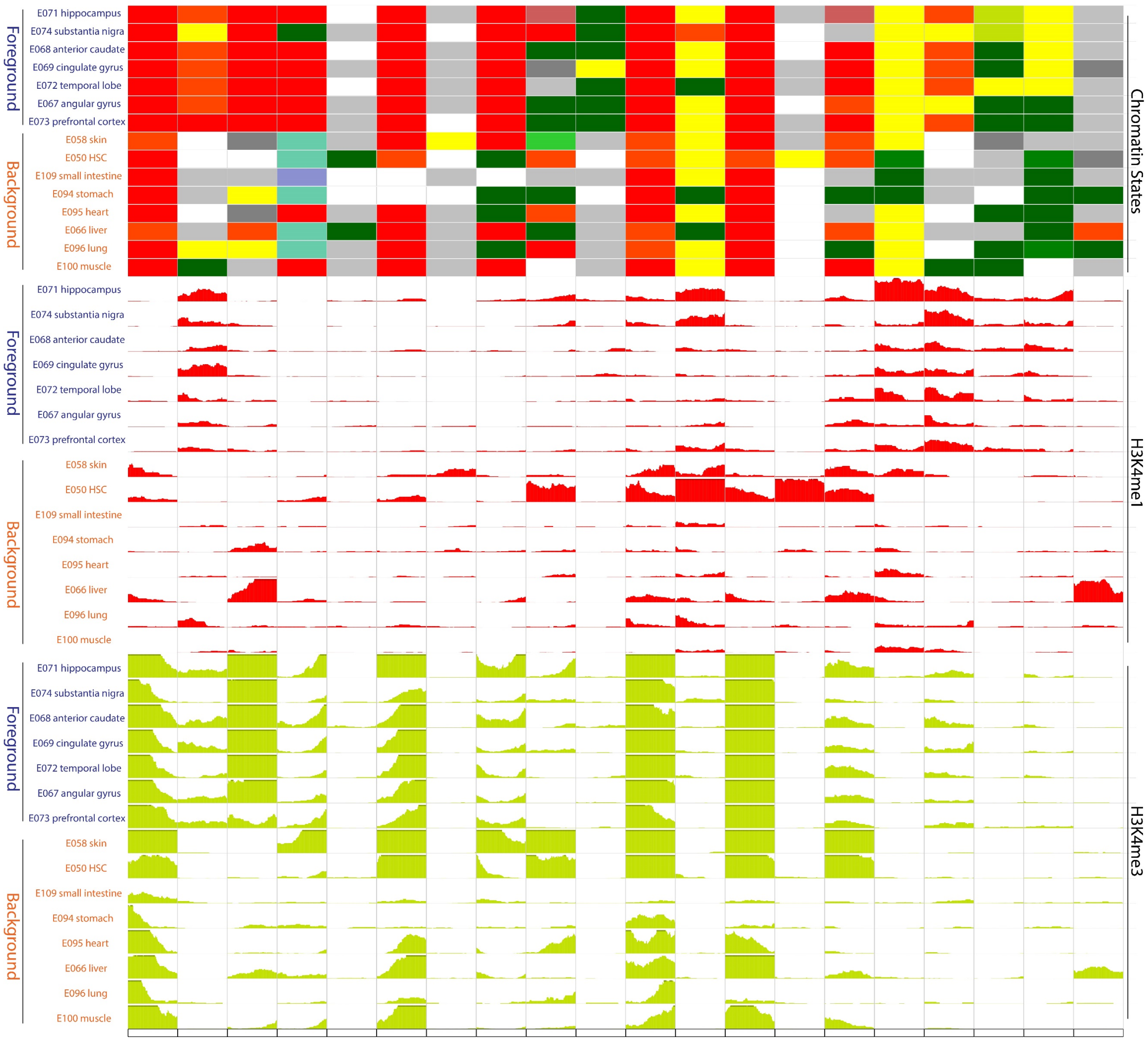


Supplementary Figure 7. Enrichment of H3K27ac peaks in brain and non-brain tissues for shared and unique brain-specific enhancers identified by Fisher's exact test and FANTOM5.



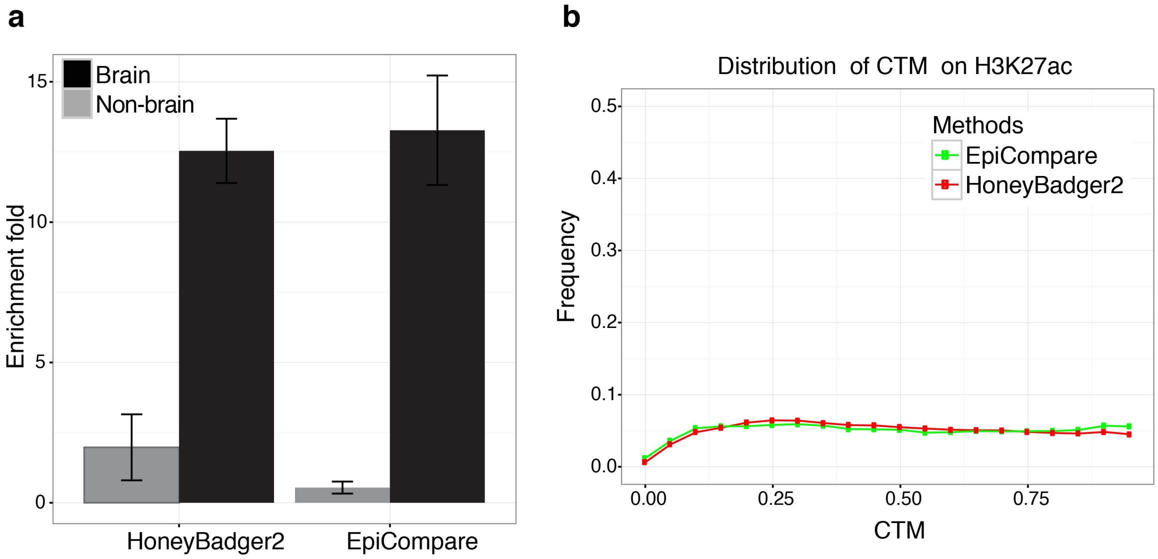
Supplementary Figure 8. Visualization of randomly picked 20 brain-specific enhancers identified from Fisher's exact test but not FANTOM5.

The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 15 state, H3K4me1 signal track (-log10(p-value)), H3K4me3 signal track (-log10(p-value)). Yellow color in ChromHMM state track represents enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Foreground samples are labeled as blue and background samples are labeled as brown.



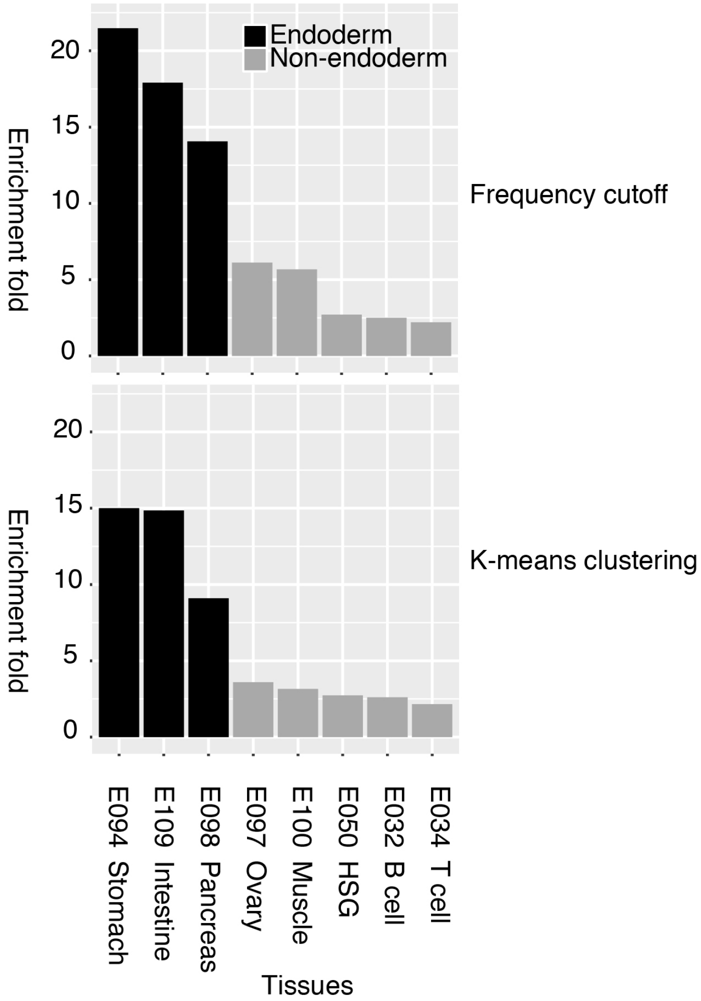
Supplementary Figure 9. Visualization of randomly picked 20 brain-specific enhancers identified from FANTOM5 but not Fisher's exact test.

The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 15 state, H3K4me1 signal track (-log10(p-value)), H3K4me3 signal track (-log10(p-value)). Yellow color in ChromHMM state track represents enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Foreground samples are labeled as blue and background samples are labeled as brown.



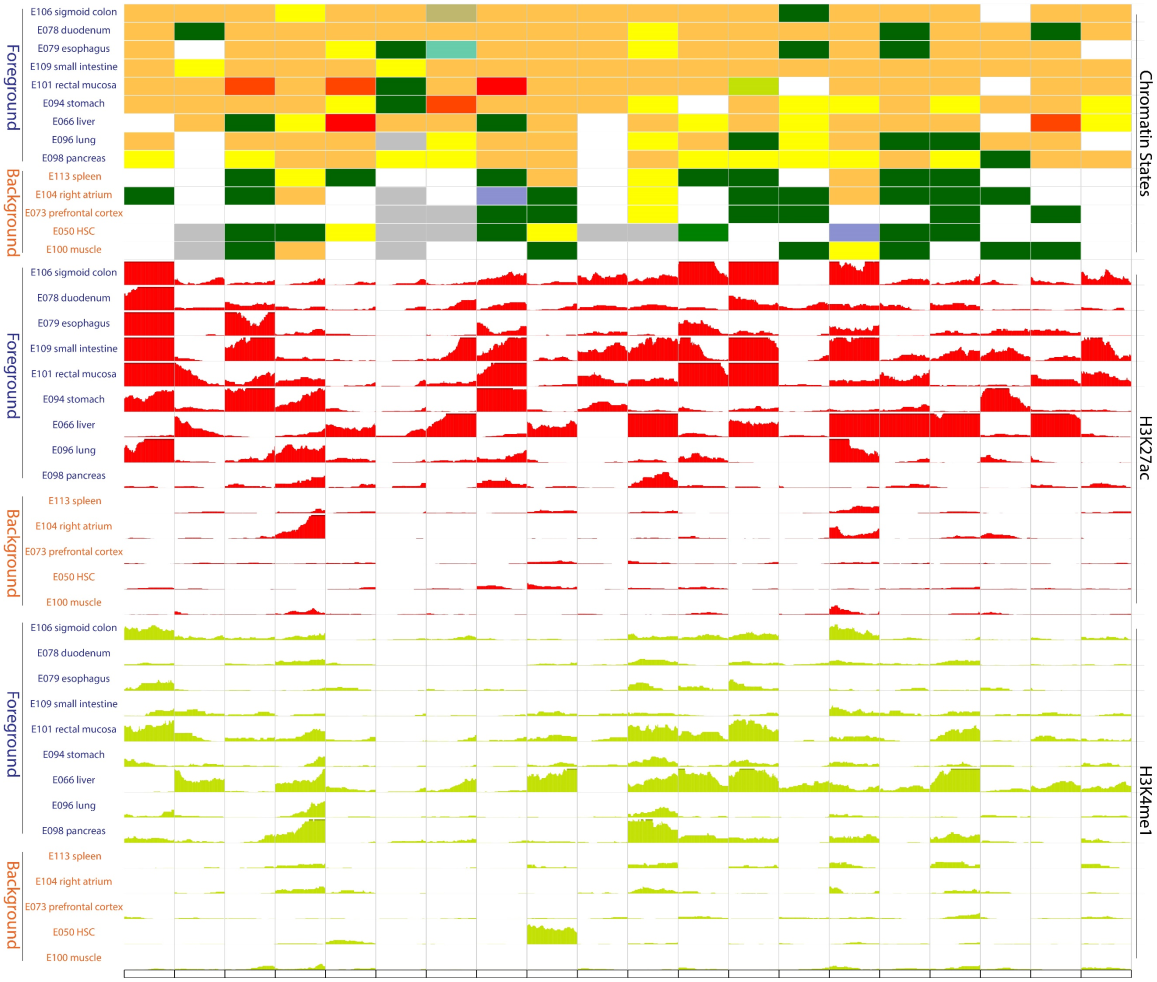
Supplementary Figure 10. Comparison of k-means clustering method in EpiCompare and HoneyBadger2 tool for identifying adult brain-specific enhancers.

(**a**) The enrichment of H3K27ac peaks in in brain and non-brain tissues for adult brain-specific enhancers identified by clustering methods in EpiCompare and HoneyBadger2. (**b**) H3K27ac RPKM-based CTM distribution in brain tissues for adult brain-specific enhancers identified by clustering methods in EpiCompare and HoneyBadger2.



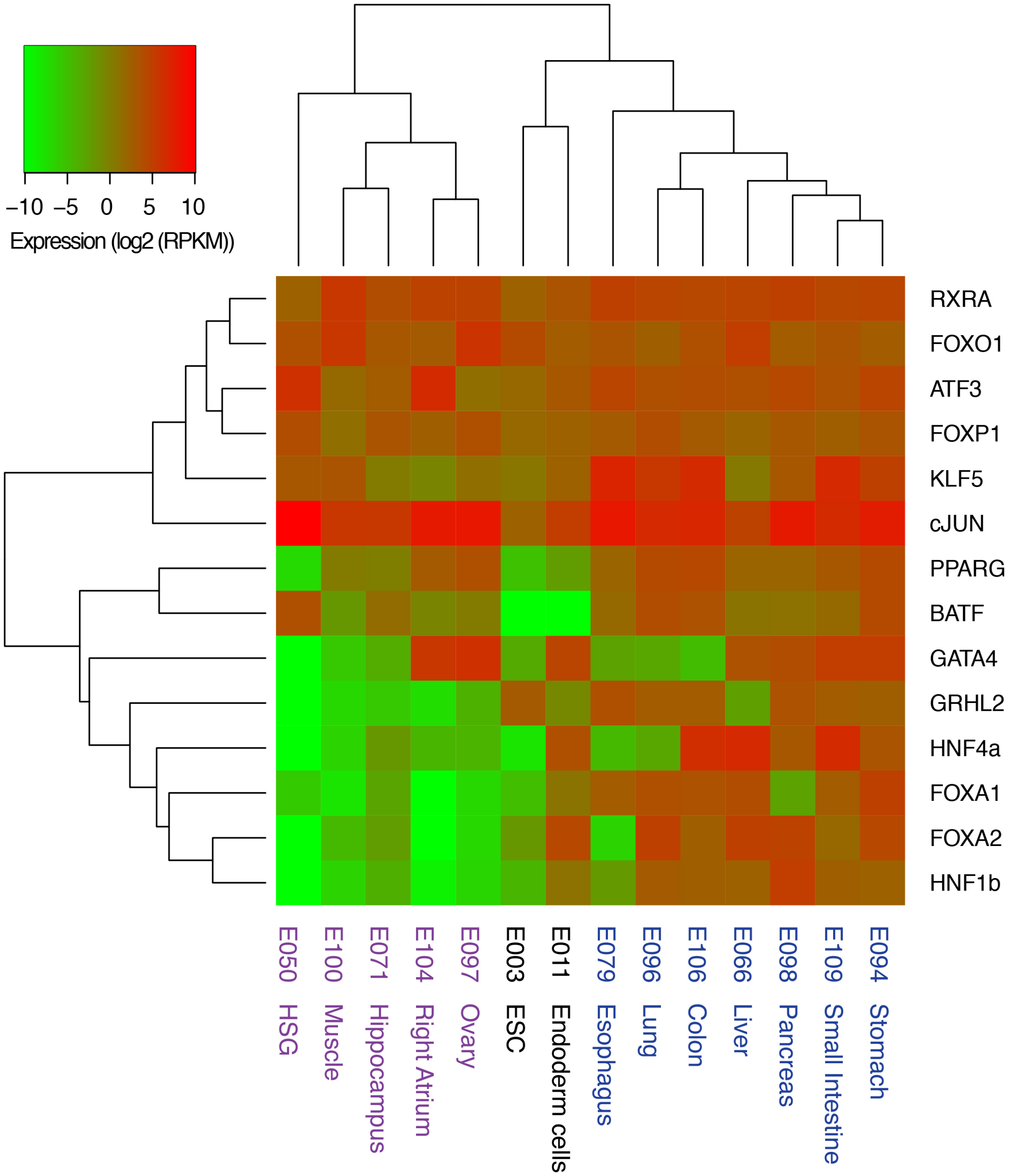
Supplementary Figure 11. Enrichment of DHS for endoderm-specific enhancers identified by frequency cutoff and k-means clustering method.

Regions identified by two methods are highly enriched for DHS in endoderm tissues comparing to other tissues.



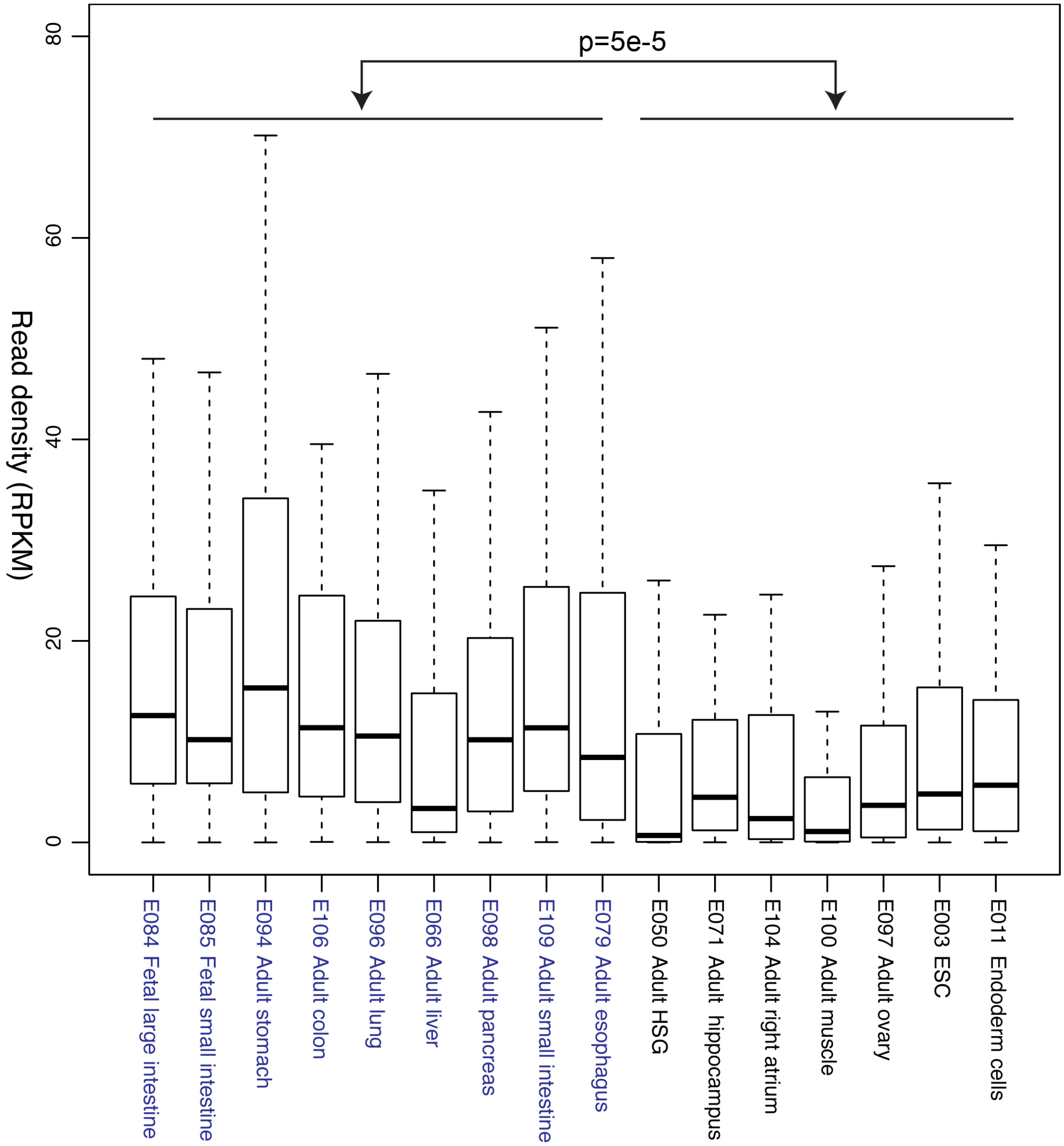
Supplementary Figure 12. Visualization of randomly picked 20 endoderm-specific enhancers identified from Fisher's exact test.

The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 18 state, H3K27ac signal track (-log10(p-value)), H3K4me1 signal track (-log10(p-value)). Yellow and orange color in ChromHMM state track represent enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state. Foreground samples are labeled as blue and background samples are labeled as brown.



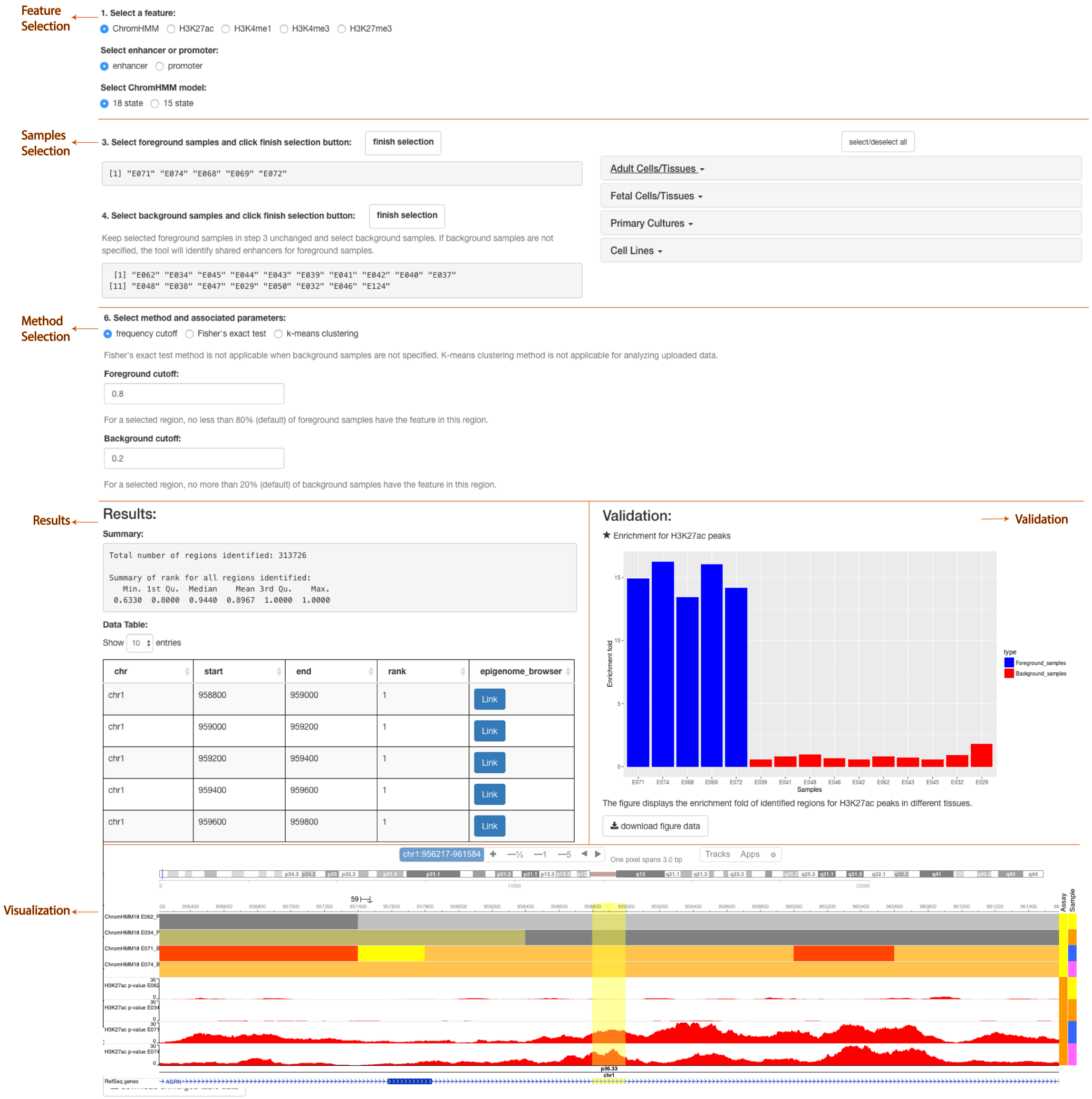
Supplementary Figure 13. The expression of enriched transcription factors in different tissues.

Endoderm tissues (blue), non-endoderm tissues (purple), ESCs and ESC-derived endoderm cells (black). The clustering is based on Euclidian distance.



Supplementary Figure 14. The distribution of expression levels of 72 predicted target genes.

Endoderm tissues (blue) and other tissues (purple). The median expression level of these genes in endoderm tissues is significantly higher than that in other tissues (t-test, p=5e-5).



1. Select a feature:
● ChromHMM ○ H3K27ac ○ H3K4me1 ○ H3K4me3 ○ H3K27me3

Select enhancer or promoter:
● enhancer ○ promoter

Select ChromHMM model:
● 18 state ○ 15 state

3. Select foreground samples and click finish selection button: finish selection

```
[1] "E071" "E074" "E068" "E069" "E072"
```

4. Select background samples and click finish selection button: finish selection

Keep selected foreground samples in step 3 unchanged and select background samples. If background samples are not specified, the tool will identify shared enhancers for foreground samples.

```
 [1] "E062" "E034" "E045" "E044" "E043" "E039" "E041" "E042" "E040" "E037"
[11] "E048" "E038" "E047" "E029" "E050" "E032" "E046" "E124"
```

6. Select method and associated parameters:
● frequency cutoff ○ Fisher's exact test ○ k-means clustering

Fisher's exact test method is not applicable when background samples are not specified. K-means clustering method is not applicable for analyzing uploaded data.

Foreground cutoff: 0.8

For a selected region, no less than 80% (default) of foreground samples have the feature in this region.

Background cutoff: 0.2

For a selected region, no more than 20% (default) of background samples have the feature in this region.

Results:

Summary:

```
Total number of regions identified: 313726

Summary of rank for all regions identified:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6330  0.8000  0.9440  0.8967  1.0000  1.0000
```

Data Table:

Show 10 entries

chr	start	end	rank	epigenome_browser
chr1	958800	959000	1	Link
chr1	959000	959200	1	Link
chr1	959200	959400	1	Link
chr1	959400	959600	1	Link
chr1	959600	959800	1	Link

Validation:

★ Enrichment for H3K27ac peaks

The figure displays the enrichment fold of identified regions for H3K27ac peaks in different tissues.

download figure data

Supplementary Figure 15. The main components of the website.

The home page includes selection of database, samples and methods. The result page includes a table of identified regions and validation analysis. The links in the table link to WashU Epigenome Browser for visualization of regions in selected samples.

Supplementary Note

1. Validation methods

We envision that most users would use our tool to define tissue or cell type-specific enhancers, therefore we provide additional validation process for this type of tasks. To validate the genomic regions identified are specific to certain tissue or cell types, we use three methods. First, GREAT analysis[83] is performed on identified regions (regions are not merged and single-nearest gene option is used) to determine if they are enriched for biological processes related to the queried tissue or cell types. The whole genome is used as background.

Second, enrichment for DNase I hypersensitive sites (DHS) and H3K27ac peaks in different tissue and cell types are calculated for identified regions, since DHS and H3K27ac modification are marks of regulatory regions. Enrichment is defined as below:

$$enrichment = \frac{\#bp \text{ in overlapped regions}/\#bp \text{ in DHS or H3K27ac peaks}}{\#bp \text{ identified regions}/\#bp \text{ in hg19 genome}}$$

Third, a tissue enrichment index for enhancers is calculated using H3K27ac RPKM (Reads Per Kilobase per Million mapped reads), since H3K27ac marks enhancers. A tissue enrichment index has been routinely used to identify tissue-specific genes[102,103]. Generally, a high tissue enrichment index represents tissue-specific regions. The tissue enrichment index is a contribution measure (CTM)[82], which is calculated as the following.

H3K27ac RPKM is calculated for N samples for each region and transformed into a vector X:

$$X = (x_1, x_2, \ldots, x_i, \ldots, x_N)$$

where x_i is the RPKM for one sample in one region. X is then converted into a cosine vector $X_{\cos\theta}$:

$$X_{\cos\theta} = (\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_i, \ldots, \cos\theta_N)$$

where $\cos\theta_i = x_i/|X|$, and $|X|$ is the magnitude of X. The tissue enrichment index in a tissue with k samples is:

$$\mathrm{CTM}=\sqrt{\sum_{i=1}^{k} \cos\theta_i{}^2}$$

Here CTM can be calculated for each tissue. The range of CTM is 0 to 1, with a higher CTM value in one tissue representing enrichment in that tissue.

2. Effect of sample size

To examine how the number of foreground samples affects the quality of regions identified, we fixed background samples to 91 non-brain samples with H3K27ac data and picked N (N=1, 2, 3, 5 or 7) samples from 7 adult brain samples as foreground samples. For each N, we chose 5 combinations of N samples selected from 7 brain samples and calculated enrichment of H3K27ac peaks in 7 brain samples for each combination. Mean and variance were calculated on the 35 enrichment values in 5 combinations. In the case of N=7, mean and variance were calculated on 7 enrichment values as there was only one combination. We set foreground cutoff and background cutoff to be 0.5 and 0.2 for frequency cutoff method, and set q-value threshold as 0.5 for Fisher’s exact test method. Other settings were set to default for all three classification methods. All enhancers were defined from 15-state ChromHMM and top 24,453 regions were used for comparison because the smallest number of regions identified in all cases was 24,453. Since k-means clustering method does not have ranks, we randomly picked 24,453 regions identified by k-means clustering method. To examine how the number of background samples affects the quality of regions identified, we fixed foreground samples to 7 adult brain samples and picked 5, 10, 30 or 91 samples from 91 non-brain samples as background samples using 5 combinations like above (**Fig. 2**).

3. Identify endoderm-specific enhancers

To test the performance of EpiCompare, we designed a simple epigenome combination comparison to identity endoderm-specific enhancers. We hypothesized that adult tissues that are derived from the endoderm should share enhancers that are specific for endoderm. Thus, we selected one sample from each adult tissue derived from the endoderm as foreground samples, including stomach, colon, liver, pancreas, lung, duodenum, esophagus, small intestine, and rectum. The background samples chosen were 34 samples in adult tissues including blood, brain, fat, heart, muscle, ovary, spleen, and aorta. We identified 13,728 regions using the frequency cutoff method (foreground cutoff = 0.7, background cutoff = 0.2), 46,859 regions using Fisher's exact test method (q-value cutoff = 0.05), and 29,386 regions using k-means clustering method with 140 clusters (default settings).

4. TF-binding motif enrichment analysis

Motif enrichment analysis was performed using the HOMER tool[85]. Enrichment for known motifs was used. The tool was also used to annotate the closest gene for each region. In the example of brain enhancers, we connected enhancers to target genes using GREAT. We confirmed that the rules for GREAT and HOMER are very similar but not identical. 90% of protein-coding genes identified from HOMER for brain-specific enhancers overlapped with that from GREAT. The analysis we performed using GREAT and/or HOMER are examples of post-EpiCompare analysis that users can perform on their own, or they can replace the analysis with other user-defined analysis on the collection of tissue or cell type-specific epigenetically marked regions returned by EpiCompare.

5. Construction of gene regulatory network

To build putative regulatory network common for adult tissues derived from the endoderm, we first identified transcription factors (TFs) that were highly expressed in endoderm tissues comparing to other tissues ($p<0.05$, t-test) and highly enriched in TF-binding motif enrichment analysis ($p<10^{-20}$). The TFs satisfying the requirements were FOXA1, FOXA2, HNF1, HNF4a. We then built a network for these TFs by linking each TF to its target genes using the nearest target gene method from the HOMER tool. Finally we identified target genes for each TF-bound endoderm-specific enhancers and used the intersection of the target genes as shared target genes for 4 TFs. These were further filtered by requiring a RPKM of >1 in at least one endoderm tissue.

6. Examine how sample size affects statistical power of Fisher's exact test

To understand how statistic power of Fisher's exact test depends on sample size, we examined how minimal q-value changes with different number of foreground samples and background samples (**Supplementary Fig. S16**). It can be seen that with increased number of foreground samples and background samples, the minimal q-value decreases greatly and therefore statistic power increases.

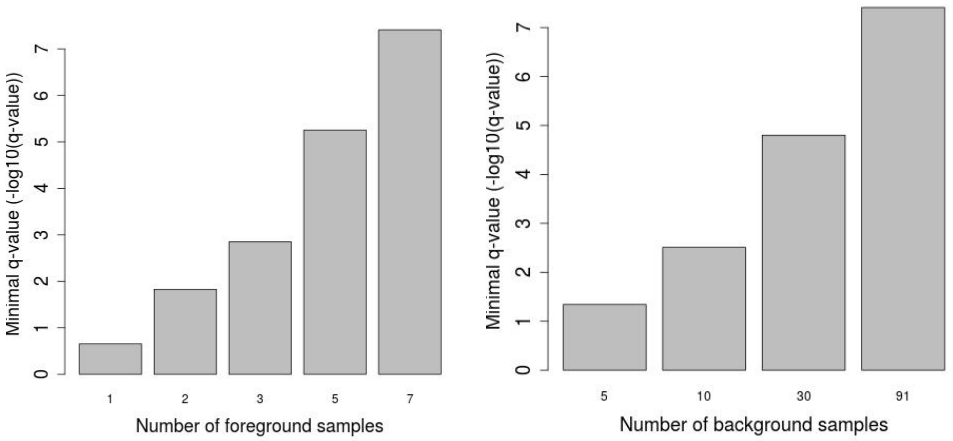


Fig. S16. How minimal q-value changes with different number of foreground samples and background samples in identifying brain-specific enhancers with Fisher's exact test.

7. Examine the false positive rate of Fisher's exact test

We examined how the number of brain-specific enhancers identified from Fisher's exact test changed with different q-value cutoff (**Supplementary Fig. S17**). To estimate the false positive rate, we randomly picked 7 samples (not including any adult brain sample) from 98 samples as foreground samples and the rest 91 as background samples and identified enhancers specific for the randomly selected 7 samples. We repeated this process 10 times. Interestingly, no enhancers were identified with q-value less than 0.1. This suggests that Fisher's exact test method along with multiple hypothesis test correction can effectively control false positive rate.

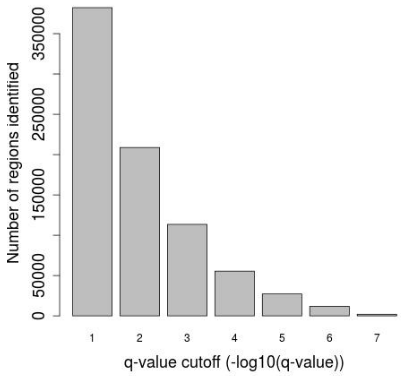


Fig. S17. How the number of regions identified with different q-value cutoff in identifying brain-specific enhancers with Fisher's exact test

8. Find optimal cluster number

we calculated sum of square errors (SSE) for a series of cluster number and identified the optimal cluster number using the elbow method[101]. To reduce the time complexity, we randomly picked 200k data points for this analysis. Here we included an example of a plot of SSE for enhancers defined by 15-state ChromHMM model (**Supplementary Fig. S18**). There was a knee point between 100 and 150. To further find the optimal number, we utilized the silhouette method, which measures how closely a point matched to data within its cluster and matched to data of the neighboring cluster. A higher silhouette value implies a more appropriate clustering. Since this method can only be done on small number of data, we randomly picked 20% of data from each cluster for this analysis. As seen from the plot, the highest value between 100 and 150 was 140 (**Supplementary Fig. S19**). Therefore, we determined the optimal cluster number to be 140 for enhancers from 15-state ChromHMM model. We used similar methods to find optimal cluster number for other features and found the optimal cluster number was close for all features and was around 140. So we provide the optimal cluster number for all features to be 140. In addition to the default number, we provide several other options (i.e., cluster number 90, 200 and 250) to give users flexibility.

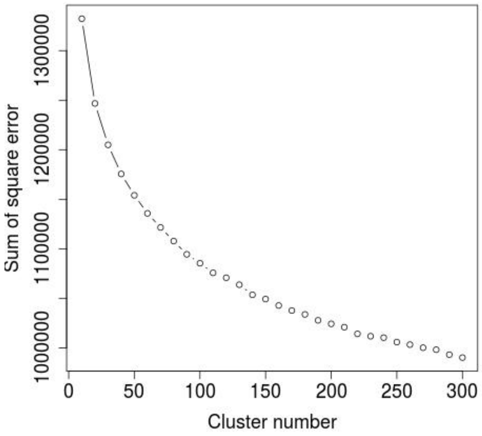


Fig. S18. The sum of square error for k-means clustering with different cluster number.

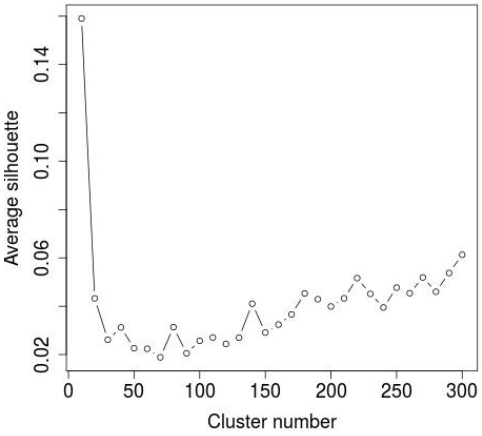


Fig. S19. The average silhouette value for k-means clustering with different cluster number.

9. Compare EpiComapre with ChromDiff

We compared identifying brain-specific enhancers using Fisher's exact test and ChromDiff on union of enhancers in seven adult brain samples. Same foreground and background samples were used for two methods. We used one-tailed Mann–Whitney–Wilcoxon test for ChromDiff.

We identified 208,804 regions using the Fisher's exact test, and 283,267 regions using ChromDiff. For fairness we kept the top 208,804 from Fisher's exact test. They overlapped by 167,691. We plotted enrichment of H3K27ac for the shared regions, as well as regions unique to each method, see **Supplementary Fig. S20.** We then randomly picked 20 regions that were unique to each method, and visualized them on the WashU Epigenome Browser in gene set view, see **Supplementary Fig. S21 and S22**. It can be seen that enhancers identified from Fisher's exact test and ChromDiff largely overlapped (80%). Enhancers that were unique to Fisher's exact test had much stronger enrichment of H3K27ac in brain samples than ChromDiff but also had higher enrichment in background samples. Therefore anecdotally EpiCompare seems to have better sensitivity, while ChromDiff seems to exhibit better specificity, at a comparable statistical cutoff. ChromDiff is a command line only program, while EpiCompare provides a much more user-friendly interface and includes access to WashU Epigenome Browser, allowing biologists to better explore their result.

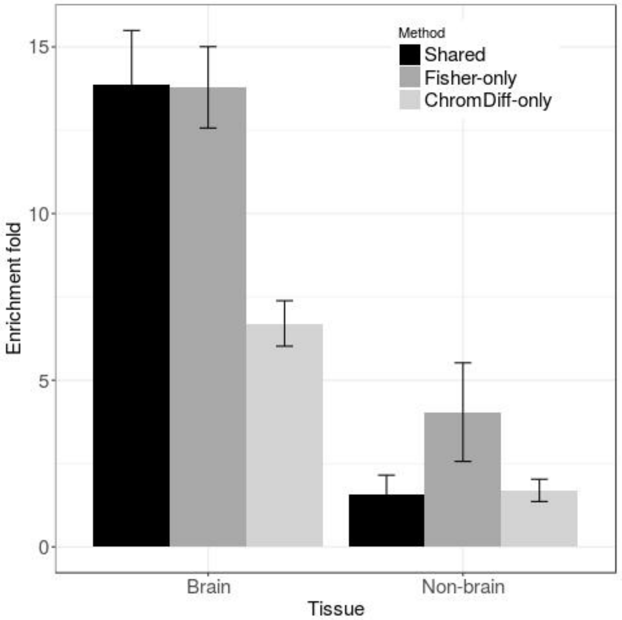


Fig. S20. Enrichment of H3K27ac peaks in brain and non-brain tissues for shared and unique brain-specific enhancers identified by Fisher's exact test and ChromDiff.

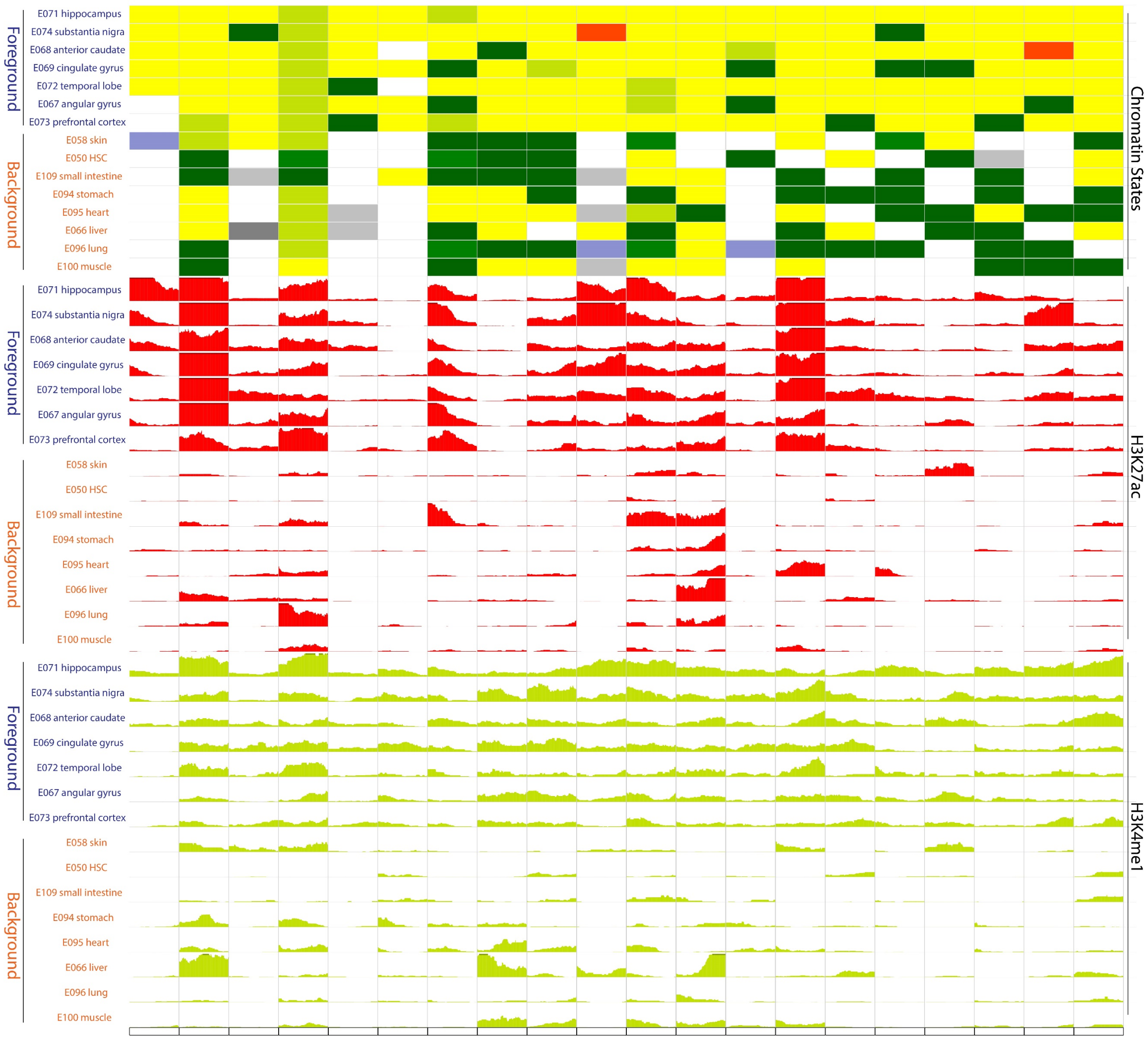


Fig. S21. Visualization of randomly picked 20 brain-specific enhancers identified from Fisher's exact test but not ChromDiff. The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 15 state, H3K27ac signal track (-log10(p-value)), H3K4me1 signal track (-log10(p-value)). Yellow color in ChromHMM state track represents enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Foreground samples are labeled as blue and background samples are labeled as brown.

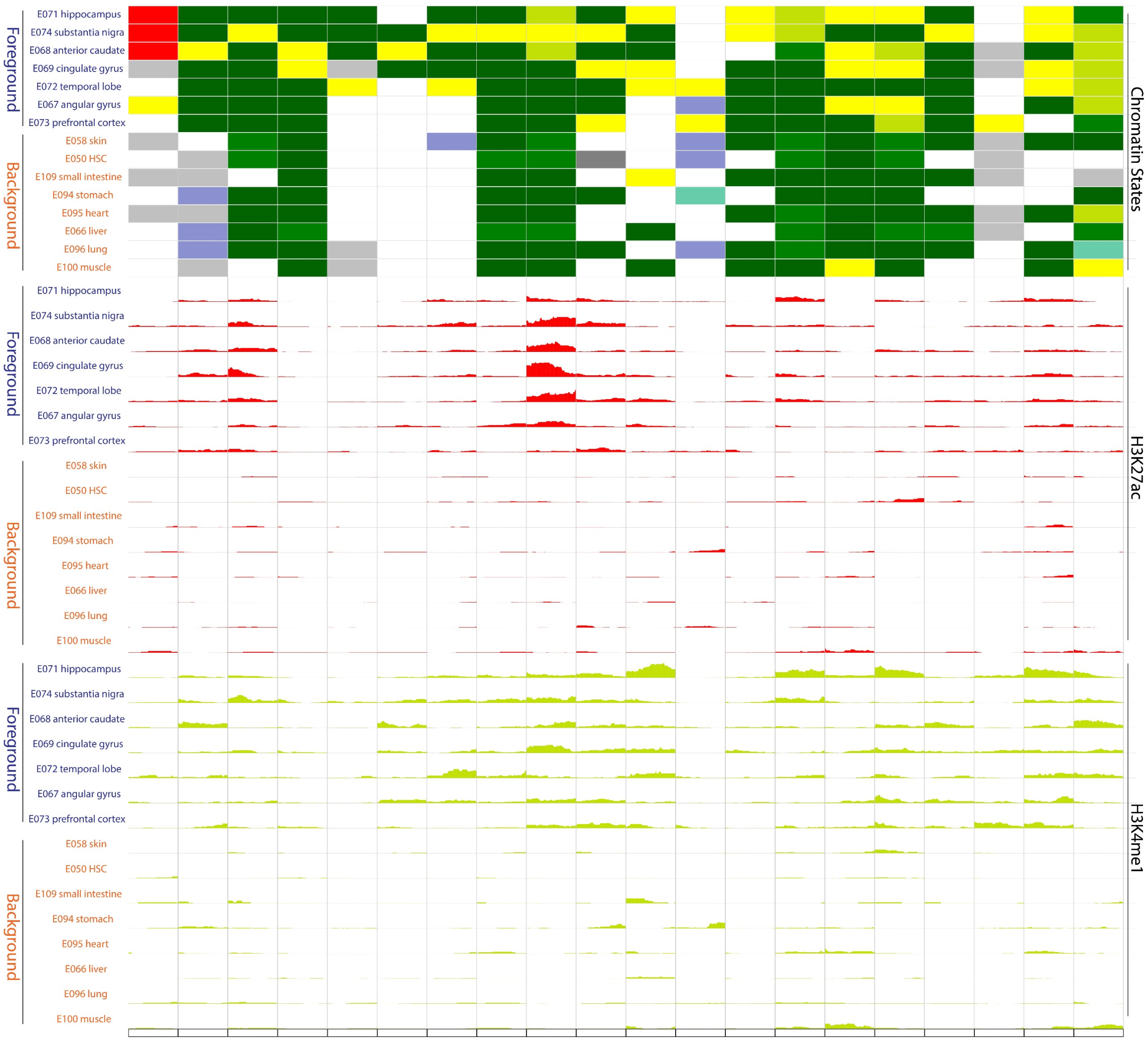


Fig. S22. Visualization of randomly picked 20 brain-specific enhancers identified from ChromDiff but not Fisher's exact test. The regions are put together using gene & region set function in WashU Epigenome Browser. 3 types of tracks are included: ChromHMM 15 state, H3K27ac signal track (-log10(p-value)), H3K4me1 signal track (-log10(p-value)). Yellow color in ChromHMM state track represents enhancer state, green represents transcribed state, white represents quiescent state and grey represents repressed state. Information about other colors can be found in http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Foreground samples are labeled as blue and background samples are labeled as brown.

Supplementary Tables

Gene	Ensembl ID
ERRFI1	ENSG00000116285
LIPH	ENSG00000163898
FARP2	ENSG00000006607
SLC45A1	ENSG00000162426
C1orf116	ENSG00000182795
ELF3	ENSG00000163435
FAM3D	ENSG00000198643
MECOM	ENSG00000085276
ATP8B1	ENSG00000081923
VEGFA	ENSG00000112715
SUSD1	ENSG00000106868
TMPRSS2	ENSG00000184012
EXT1	ENSG00000182197
PRLR	ENSG00000113494
RHPN2	ENSG00000131941
SERINC2	ENSG00000168528
KIAA1324	ENSG00000116299
BMP2	ENSG00000125845
ACVR2A	ENSG00000121989
MYO5C	ENSG00000128833
MYO10	ENSG00000145555
CHN2	ENSG00000106069
PIGR	ENSG00000162896
ROR1	ENSG00000185483
IFFO2	ENSG00000169991
TJP2	ENSG00000119139
ANKRD40	ENSG00000154945
PARVA	ENSG00000197702
NRP2	ENSG00000118257
DHRS2	ENSG00000100867

SLC22A23	ENSG00000137266
FOXQ1	ENSG00000164379
IER3	ENSG00000137331
WWC1	ENSG00000113645
C14orf2	ENSG00000156411
HMGCS2	ENSG00000134240
CD55	ENSG00000196352
CCDC58	ENSG00000160124
COL21A1	ENSG00000124749
DNPEP	ENSG00000123992
KCNK5	ENSG00000164626
GPR37L1	ENSG00000170075
CAMTA1	ENSG00000171735
FZD5	ENSG00000163251
HES1	ENSG00000114315
SYF2	ENSG00000117614
TTLL6	ENSG00000170703
STK24	ENSG00000102572
F3	ENSG00000117525
CHD1L	ENSG00000131778
AMBP	ENSG00000106927
PLEKHA7	ENSG00000166689
INO80D	ENSG00000114933
NDRG1	ENSG00000104419
SH3YL1	ENSG00000035115
C9orf152	ENSG00000188959
IHH	ENSG00000163501
TTC39A	ENSG00000085831
TOX3	ENSG00000103460
PANK3	ENSG00000120137
HEXB	ENSG00000049860
AGAP1	ENSG00000157985
ABCC4	ENSG00000125257

GPR160	ENSG00000173890
KALRN	ENSG00000160145
C6orf132	ENSG00000188112
PPP2R5A	ENSG00000066027
PDXK	ENSG00000160209
ADGRV1	ENSG00000164199
LINC01549	ENSG00000232560
SPIDR	ENSG00000164808
FAM86B3P	ENSG00000173295

Supplementary Table 1. The intersection of target genes for each TF (FOXA1, FOXA2, HNF1a, HNF4a)-bound endoderm-specific enhancers.

Pathway	Raw p-value
WNT5A-dependent internalization of FZD2, FZD5 and ROR2	0.000165
VEGF and VEGFR signaling network	0.000734
BMP Signalling Pathway	0.002443
Canonical Wnt signaling pathway	0.003018
Signaling by Hippo	0.003018
Signaling by BMP	0.003327
Validated targets of C-MYC transcriptional repression	0.003471
Mesodermal Commitment Pathway	0.003743
Class B/2 (Secretin family receptors)	0.005468
VEGFR1 specific signals	0.006298
Signal Transduction	0.008242
Developmental Biology	0.009417

Supplementary Table 2. Pathway enrichment from ConsensusPathDB for 72 target genes

Chapter 3: DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods

Yu He, Hyo Sik Jang, Xiaoyun Xing, Daofeng Li, Michael J. Vasek, Joseph D. Dougherty, Ting Wang

From:

DeepH&M: Estimating single-CpG hydroxymethylation and methylation levels from enrichment and restriction enzyme sequencing methods.

Under review at Science Advances, Apr 2020

3.1 Abstract

Increased appreciation of 5-hydroxymethylation (5hmC) as a stable epigenetic mark, which defines cell identity and disease progress, has engendered a need for cost-effective, but high-resolution 5hmC mapping technology. Current enrichment-based technologies provide cheap, but low-resolution and relative enrichment of 5hmC levels while single base-resolution methods can be prohibitively expensive to scale up to large experiments. To address this problem, we developed a deep learning-based method “DeepH&M”, which integrates enrichment and restriction enzyme sequencing methods to simultaneously estimate absolute hydroxymethylation and methylation levels at single CpG resolution. Using 7-week-old mouse cerebellum data for training DeepH&M model, we demonstrated that the 5hmC and 5mC levels predicted by DeepH&M were in high concordance with whole genome bisulfite-based approaches. The DeepH&M model can be applied to 7-week old frontal cortex and 79-week cerebellum revealing the robust generalizability of this method to other tissues from various biological time points.

3.2 Introduction

A single genome can derive phenotypically unique cell types through various epigenetic modifications that instruct specific gene expression patterns[104,105]. DNA modifications, such as methylation of 5 position of cytosines (5mC) at CpG dinucleotide context, play a vital role in gene regulation, genomic imprinting, X-chromosome inactivation, and repression of transposable elements[30–33]. The recent discovery that Ten-Eleven Translocation (TET) oxidase proteins can oxidize 5mC to 5-hydroxymethylcytosine (5hmC) has spurred an effort at characterizing the landscape of 5hmC in normal and diseased tissues and deciphering its potential functional role in gene regulation[39,106–110]. Genome-wide profiling of 5hmC has found that 5hmC is not just an intermediate product of the active DNA demethylation process, but also a stable epigenetic mark correlated with gene expression. 5hmC abundance varies significantly across different tissues[111]. 5hmC is present as high as 40% of 5mC levels in Purkinje neurons[37], 5% of 5mC levels in embryonic stem cells[43], and is low (less than 1% of 5mC level) in other cell types[47]. 5hmC is enriched in promoters, gene bodies and enhancers; 5hmC levels in promoters and gene bodies are positively correlated with gene expression[45–47]. 5hmC levels in enhancers are often cell-type specific and are positively correlated with active enhancer histone marks, such as H3K4me1 and H3K27ac[48]. However, the molecular mechanism by which 5hmC might regulate the genome has yet to be fully elucidated[112].

Rapid technological innovations for mapping 5mC have cemented 5mC as a crucial epigenetic mark for cell fate. Technologies for mapping 5mC include bisulfite conversion of unmethylated cytosine to uracil, such as whole genome bisulfite sequencing (WGBS), enrichment of methylated DNA using methyl-cytosine-specific antibodies, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq), and enrichment of unmethylated regions using methylation-sensitive restriction enzymes, such as methylation-sensitive restriction enzyme sequencing (MRE-seq)[63]. The gold standard method WGBS can measure methylation genome-wide at single-base resolution but requires high coverage of the genome (at least 10x coverage for each cytosine) and therefore can be 10 times more expensive than enrichment or restriction enzyme sequencing methods[64]. MeDIP-seq enriches for methylated regions but has low resolution (~150bp)[65,66]. MRE-seq provides CpG resolution, but can only interrogate methylation status at restriction enzyme sites (~30% of the genome)[65].

Similarly, 5hmC profiling technologies advanced from immunoprecipitation/enrichment-based methods to whole genome single-base resolution. Because WGBS cannot distinguish 5hmC from 5mC, Yu et. al. developed a method called TET-assisted bisulfite sequencing (TAB-seq), where 5hmCs are first protected by glucosylation and then 5mC is completely oxidized to

5caC with TET enzyme[46]. The following bisulfite treatment can reveal which CpGs are protected and infer hydroxymethylation levels. TAB-seq can measure genome-wide 5hmC at single-base resolution but requires very high coverage to confidently call 5hmC at all cytosines. For example, for 5% 5hmC, based on binomial test with a probability of 2.22% for 5mC non-conversion rate, a coverage of 120 is required to call 5hmC at 95% confidence level (see Materials and Methods). The study from Yu, et al. could only confidently call 20% or higher 5hmC at an average coverage of 27. Often in TAB-seq experiments, both WGBS and TAB-seq libraries are deeply sequenced to parse out 5mC and 5hmC levels in a single sample. Achieving high confidence, single base resolution of 5hmC can be a heavy financial strain for large experimental designs due to the necessary sequencing depth. Therefore, many adopted the cheaper alternative of utilizing antibody-based enrichment method, such as hydroxymethylated DNA immunoprecipitation sequencing (hMeDIP-seq), which can reveal hydroxymethylated regions with limited sensitivity[45]. hMeDIP-seq can also provide relative hydroxymethylation over controls, but at the cost of low resolution. Similar to antibody-based enrichment method such as hMeDIP-seq, hmC-Seal chemically tags hydroxymethylated cytosine and enriches hydroxymethylated regions by pulling down tagged 5hmC[47,48]. hmC-Seal can pull down regions with extremely low 5hmc content and thus have higher sensitivity than hMeDIP-seq.

Because of the high cost of single-base-resolution profiling methods for 5hmC and 5mC, several computational methods were developed to estimate 5hmC and 5mC at single-base resolution. Xiao et.al. developed a random forest regression-based method MeSiC to estimate single-CpG 5mC from MeDIP-seq data[67]. Stevens et. al. took advantage of the complementary properties of MeDIP-seq and MRE-seq and developed a conditional random field-based algorithm methylCRF to effectively predict single-CpG 5mC from MeDIP-seq and MRE-seq data[68]. However, the two aforementioned algorithms cannot predict 5hmC levels. Pavlovic et al. developed a SVM/random forest-based method DIRECTION to predict single-CpG 5mC or

5hmC from histone modification and transcription factor ChIP-seq data[69]. This method can only predict binary values, either high or low 5mC/5hmC, but not the absolute quantitative level. To address these limitations, we developed a deep learning-based method DeepH&M, which integrates enrichment and restriction enzyme sequencing methods to estimate absolute single-CpG resolution hydroxymethylation and methylation levels simultaneously.

3.3 Results

3.3.1 Description of DeepH&M model

To estimate single-CpG hydroxymethylation and methylation, we developed a deep learning-based algorithm DeepH&M to integrate MeDIP-seq, MRE-seq and hmC-Seal data (**Fig. 1A**). The core of DeepH&M is to model the relationship between MeDIP-seq/MRE-seq/hmC-Seal data and TAB-seq/WGBS data using deep learning networks. The relationship between MeDIP-seq/MRE-seq data and WGBS data was well characterized previously in a conditional random filed-based algorithm, methylCRF, which was used to integrate MeDIP-seq and MRE-seq data to predict absolute methylation levels at single-CpG resolution[68]. hmC-Seal data is positively correlated with TAB-seq data while MeDIP-seq and MRE-seq data present a complex relationship with TAB-seq data (**fig. S1A**). DeepH&M model is composed of 3 modules: a regular neural network-based CpG module, a convolutional neural network-based DNA module and a regular neural network-based joint module (**Fig. 1B**). The inputs for CpG module are genomic features and methylation features (table S1) for each CpG. Genomic features include GC percent, CpG density and distance to nearest CpG island. Methylation features include MeDIP-seq, MRE-seq and hmC-Seal signal. Because CpG in proximity tends to have similar 5hmC and 5mC level (**fig. S1B**), we also include average signal for above features in neighboring windows around the target CpG. DNA module takes DNA sequence around a CpG as inputs and uses convolutional neural network to extract information from DNA sequence. The joint module combines outputs from CpG module and DNA module and predicts 5hmC and 5mC levels simultaneously.

3.3.2 Benchmarking DeepH&M model

To examine the performance of DeepH&M, we generated WGBS, TAB-seq, MeDIP-seq, MRE-seq and hmC-Seal data for 7-week-old mouse cerebellum and trained DeepH&M model with these datasets. Because DeepH&M requires 5hmC and 5mC as the labels, we used a statistical method MLML[113] to integrate TAB-seq and WGBS data to get consistent 5hmC, 5mC and total methylation. MLML can prevent obtaining negative 5mC values by subtracting TAB-seq data directly from WGBS data, and also prevent the contradiction of TAB-seq and WGBS data at some CpG sites. As a reference, we called 5hmC, 5mC and total methylation derived from MLML as "gold standard" data and evaluated our predictions against them. However, we recognize that even the gold standard data might not represent the true hydroxymethylation and methylation levels of a sample due to intrinsic limitations of profiling methods as described previously[114,115].

Our predicted 5hmC, 5mC and total methylation levels are in high concordance with gold standard results. DeepH&M recapitulates the distribution of gold standard 5hmC, 5mC, and total methylation (**Fig. 2A** and **B**). The genome-wide correlation across our predictions and gold standard data for 5hmC, 5mC and total methylation is 0.8, 0.85 and 0.85 respectively (**Fig. 2A**). Using a previously developed concordance metric (defined as the percent of CpGs with a methylation proportion difference less than 0.1 or 0.25)[116], 5hmC predictions are 86% concordant with gold standard data within 0.1 difference, 5mC predictions are 90% concordant within 0.25 difference and total methylation predictions are 91% concordant within 0.25 difference. To examine if the concordance is high only at particular 5hmC/5mC/total methylation levels, we examined the concordance at differing 5hmC/5mC/total methylation windows (**Fig.**

2C). 5hmC concordance is over 80% for 5hmC levels less than 0.4 and 45% for 5hmC levels higher than 0.4. We report that less than 1% of the CpGs in mouse cerebellum have 5hmC levels higher than 0.4. One explanation for the low concordance could be due paucity of high hmC CpGs in the training set (2 million CpGs) thus DeepH&M might have difficultly learning the rules for high 5hmC CpGs. The concordance for 5mC is relatively lower for 5mC at 0.2-0.4 window and the concordance for total methylation is low for total methylation at 0.2-0.6 window. This may be due to the difficulty in predicting intermediate methylation as the problem also existed in predictions by methylCRF[68]. The high concordance can be appreciated in the WashU Epigenome browser view of the Slc22a17 and Efs locus, where 5hmC, 5mC and total methylation levels of predicted and gold standard data are visualized (**Fig. 2D**). Furthermore, as a positive control for evaluating our predictions against gold standard data, we examined the concordance of two 7-week-old cerebellum replicates (**fig. S2**). The genome-wide correlation for 5hmC, 5mC and total methylation between two replicates is 0.82, 0.89, and 0.91 respectively, and the concordance is 88%, 92%, and 94% respectively. The concordance of our predictions with gold standard data is very close to the concordance of two replicates. These results confirm DeepH&M can estimate single-CpG hydroxymethylation and methylation with high accuracy.

Since it has been shown that 5hmC is enriched at enhancers and 5hmC levels at gene body are positively correlated with gene expression[45–47], we investigated if our 5hmC predictions can reveal these relationships. To examine the enrichment of 5hmC in genomic features, we divided CpGs into four categories based on their 5hmC levels and calculated the enrichment fold of the four CpG categories in genomic features. We found the enrichment of DeepH&M predicted 5hmC in genomic features was similar to that of gold standard 5hmC (**fig. S3A**). CpGs with high 5hmC levels by predictions or gold standard data were highly enriched for enhancers and depleted for promoters. To examine the relationship between 5hmC and gene expression, we grouped genes into four categories based on expression levels and profiled average 5hmC levels at gene body of the four categories of genes. We observed that similar to the gold standard 5hmC, the predicted 5hmC levels were positively correlated with gene expression (**fig. S3B**).

3.3.3 Factors affecting DeepH&M performance

Next, we wanted to investigate factors that may affect DeepH&M's performance. First, we examined DeepH&M's performance across different genomic features, as DNA methylation and hydroxymethylation were known to be highly non-random across the genome. The concordance is over 93% at CpG islands and promoters for 5hmC and 5mC (**Fig. 3A**). The concordance for other genomic features is over 80% for 5hmC and over 87% for 5mC. Because most CpG islands (CGIs) are lowly methylated and only a small portion of CGIs are highly methylated, we wanted to see if DeepH&M can distinguish highly methylated CGIs from lowly methylated CGIs. We divided CGIs into lowly methylated CGI and highly methylated CGIs based on total methylation levels, and then examined the concordance of predictions and gold standard data in these two types of CGIs. At lowly methylated CGIs, the concordance for 5hmC and 5mC are 99.9% and 99.8%, respectively (**Fig. 3B**). At highly methylated CGIs, the concordance for 5hmC and 5mC are 95% and 98%. These results indicate that DeepH&M's predictions are determined by experimental data instead of a learned assumption that all CGIs are lowly methylated. Second, since the accuracy of methylation levels from TAB-seq and WGBS data are significantly influenced by sequencing coverage, we examined DeepH&M's performance across differing CpG coverage from TAB-seq and WGBS data. The concordance for 5hmC and 5mC increases steadily from less than 10x coverage to over 10x coverage (85% to 88% for 5hmC, 78% to 89% for 5mC) (**Fig. 3C**). Thus, the lower concordance at lower coverage is likely a consequence of lower confidence in gold standard data, underscoring the robustness of our algorithm. Third, we examined DeepH&M's performance across regions with differing CpG density, as CpG density is a confounding factor for our enrichment-based sequencing methods, MeDIP-seq and hmC-Seal, which do not work optimally for regions with low CpG density. Indeed, we observed increasing concordance for 5hmC and 5mC with increasing CpG density. Note that the concordance was greater than 0.8 even at lowest CpG density; it increased to over 88% (5hmC) and 92% (5mC) for high CpG density regions that most of the current investigations focus on (**Fig. 3D**).

3.3.4 Generalizability of DeepH&M model to explore hydroxymethylation and methylation dynamics

Finally, we wanted to test whether DeepH&M model, trained on data from 7-week-old mouse cerebellum, can be generalized to data of other samples. This includes whether DeepH&M can predict differentially hydroxymethylated regions (DHMR) and differentially methylated regions (DMR) between two samples. We generated WGBS, TAB-seq, MeDIP-seq, MRE-seq and hmC-Seal data for 79-week-old mouse cerebellum as we wanted to explore 5hmC changes during aging. Using DeepH&M model from 7-week-old mouse cerebellum, we predicted 5hmC and 5mC for 79-week-old mouse cerebellum. We performed similar concordance analysis between predictions and gold standard data for 79-week-old mouse cerebellum. The overall performance of DeepH&M model in 79-week-old mouse cerebellum is similarly high as 7-week-old mouse cerebellum (**Fig. 4**, **A** to **C**). The genome-wide correlation for 5hmC, 5mC and total methylation between predictions and gold standard data is 0.81, 0.86, and 0.86 respectively, and the concordance is 84%, 91%, and 92% respectively. As illustrated by the WashU Epigenome browser view, there is high concordance between DeepH&M prediction and gold standard data across 5hmC, 5mC and total methylation levels in the 5'UTR and first exon of Kcnd2 gene (**Fig. 4D**).

Recent research suggests that epigenetic mechanisms, DNA methylation in particular, play a central role in the aging process[117]. Using antibody-based methods to quantify 5hmC levels, several studies reported global levels of 5hmC increase significantly in mouse cerebellum during aging, but remain stable in mouse hippocampus[118,119]. Furthermore, a recent study used single-base-resolution sequencing method (oxBS-seq) to measure 5hmC at single sites in mouse hippocampus and found no global 5hmC changes[120]. However, due to low sequencing depth (2X), the study only examined 5hmC changes at chromosome level and genomic element level, such as CGIs and promoters, and could not provide single-base resolution 5hmC dynamics at local regions.

In this study, we explored whether DeepH&M could reveal how 5hmC changes globally and locally in mouse cerebellum during aging. We report that global 5hmC levels increase by 20% from 7 weeks to 79 weeks and that global 5mC levels do not change (**table S2**). Next, we examined if there are 5hmC and 5mC changes in specific regions during aging by calling differentially hydroxymethylated regions (DHMRs) and differentially methylated regions (DMRs). First, we identified 524 DHMRs between hmC-Seal data of 7-week-old and 79-week-old mouse cerebellum using DiffBind[121]. We wanted to see if 5hmC changes in these DHMRs are similar between predictions and gold standard data. Indeed, the hyperDHMRs have significantly higher 5hmC in both gold standard data and predictions, and hypoDHMRs have significantly lower 5hmC in both gold standard data and predictions (**Fig. 5A**). Thus, both gold standard data and DeepH&M predictions support DHMRs defined by hmC-Seal data. Second, we defined DHMRs and DMRs by comparing TAB-seq and WGBS data between 7-week-old and 79-week-old cerebellum using the tool DSS[122]. We examined whether these DHMRs/DMRs are supported by DeepH&M data. Indeed, the differences predicted by DeepH&M are highly significant, and they are concordant with differences defined by gold standard data, although the overall magnitude tends to be smaller (**Fig. 5**, **B** and **C**).

We also examined enrichment of biological processes for these DHMRs and DMRs using GREAT[83]. We report that hyperDHMRs are enriched near genes that regulate synaptic plasticity and transporter activity (**fig. S4A**) and that hyperDMRs are enriched in genes responsible for neuron axonogenesis (**fig. S4B**). There were no significantly enriched terms associated with hypoDMRs and hypoDHMRs, possibly due to the small number of hypoDMRs and hypoDHMRs. As an example, **Fig. 4D** illustrates one of the numerous differentially hydroxymethylated regions between 7-week-old and 79-week-old cerebellum. The 5hmC changes at this region is supported by changes of gold standard 5hmC, predicted 5hmC and also hmC-Seal signal between the two ages. These results suggest that DeepH&M can predict DHMRs and DMRs between two samples.

The above analysis demonstrates that DeepH&M model, trained on data from 7-week-old mouse cerebellum, can be generalized to 79-week-old mouse cerebellum. We wanted to examine whether our DeepH&M model can be also generalized to 7-week-old mouse cortex as 5hmC levels in cortex is much higher than cerebellum. We found the overall performance of DeepH&M model for 5hmC is a little lower in cortex than in cerebellum (concordance: 72% vs 86%), and the performance for 5mC and total methylation is similar to cerebellum (**Fig. 6**, **A** to **C**). The genome-wide correlation for 5hmC, 5mC and total methylation between predictions and gold standard data is 0.65, 0.82, and 0.89 respectively, and the concordance is 72%, 89%, and 92% respectively. We can see that 5hmC distribution in cortex is distinct from cerebellum (**Fig. 2B** vs. **Fig. 6B**) and the mean 5hmC level in cortex is almost twice as high as in cerebellum (0.19 vs 0.11). Satisfyingly, DeepH&M can still recapitulate the distribution of gold standard 5hmC and 5mC and total methylation. These results suggest DeepH&M model trained from cerebellum is not only generalizable to other cerebellum samples at different ages, but also generalizable to adult frontal cortex. We also applied our DeepH&M model to mouse fetal cortex which has much lower global 5hmC levels than adult cortex. The genome-wide correlation for

5hmC, 5mC and total methylation between predictions and gold standard data provided in Lister et al.[123] is 0.44, 0.63, and 0.65 respectively, and the concordance is 61%, 84%, and 94% respectively (**fig. S5**). The extremely low concordance for 5hmC in fetal cortex may be explained by the rather big global differences in 5hmC distribution in adult and fetal cortex.

3.4 Discussion and Conclusions

5hmC is known to be an intermediate, but stable, epigenetic feature of the active DNA demethylation process. However, the molecular mechanisms underlying the role of 5hmC in gene regulation remains largely unknown. Furthermore, the loss of 5hmC has been identified as a hallmark of most types of human cancers. Many cancers are characterized by down-regulation of or deleterious mutations in TET or isocitrate dehydrogenase IDH1/IDH2 (co-factors of TET enzymes) genes, which reduces the rate of oxidization of 5mC into 5hmC[106,108,109]. It's important to note that many of these studies employ hMeDIP-seq technology to profile tumor and matched-tumor samples, therefore, there is a lack of high-resolution hydroxymethylomes of tumors.

Understanding the mechanisms underlying 5hmC's roles in development and tumorigenesis can benefit from profiling 5hmC levels at genome-wide, single-base resolution. As shown in Wen et al.[124], high-resolution 5hmC profiling of the human brain revealed novel 5hmC signatures, such as high hydroxymethylation levels near 5'splicing sites and transcription-correlated hmC levels on the sense strand of the gene, that hMeDIP-seq would not be able to detect due to inherent limitations of the technology. Identifying these novel signatures could hold the key in deciphering the biological machineries that 5hmC could potentiate. Currently, TAB-seq and oxidative-bisulfite sequencing (oxBS-seq) are the gold standard methods for providing single-CpG-resolution DNA hydroxymethylomes[46,125]. These two methods require very high coverage to confidently call 5hmC at all cytosines. The coverage required for oxBS-seq is even higher due to the fact that oxBS-seq measures 5hmC indirectly through subtracting measured 5mC from measured total methylation. The high cost associated with the high coverage is a significant barrier for individual laboratories to adopt TAB-seq and oxBS-seq as a routine assay for DNA hydroxymethylomes. Indeed, so far only a few cell types have deeply sequenced hydroxymethylomes at single-base resolution[46,115,123,124,126–130].

To overcome this potential cost-barrier problem, we have developed a deep learning-based algorithm DeepH&M, which integrates enrichment and restriction enzyme sequencing methods to estimate the absolute levels of hydroxymethylation and methylation at single CpG resolution. The cost of the three assays combined is <5% of WGBS and TAB-seq. About 50-100 million MeDIP reads, 30 million MRE reads and 50 million hmC-Seal-seq reads are sufficient for measuring a hydroxymethylome with DeepH&M, which translates to roughly 3x coverage of the human or mouse genome. Also, TAB-seq requires ~3ug of genomic DNA while MeDIP-seq, MRE-seq, and hmC-Seal can be generated from 100ng or less input thus allowing DeepH&M to be more amenable to rare or difficult-to-procure cells or samples. Compared to 100x coverage for TAB-seq and 20x coverage for WGBS, our method can minimize the cost of generating a complete hydroxymethylome by 40-fold. Furthermore, DeepH&M can estimate for all CpGs while WGBS and TAB-seq miss a significant fraction of the genome due to low coverage. As mentioned previously, previous TAB-seq study on H1 cells could only confidently call 20% or higher 5hmC at a coverage of 27 and thus identified less than 1 million hydroxymethylated CpGs[46].

One caveat to DeepH&M is that TAB-seq and WGBS libraries must be sequenced initially to generate training data for the cell type of interest. Since creating comprehensive hydroxymethylome and methylome can be cost-prohibitive, we explored alternative methods of generating training data. Currently, Infinium MethylationEPIC BeadChip Kit (Illumina, WG-317-1001) can profile the methylation levels from roughly 850,000 CpGs at single-nucleotide resolution for human. To address whether methylation microarray results could be utilized as training set, we asked whether DeepH&M can be trained on 850,000 CpGs in our mouse data. Compared to 2 million CpG training data, which has 86% 5hmC and 90% 5mC concordance, DeepH&M can still predict with 83% and 89% concordance for 5hmC and 5mC respectively. Therefore, to reduce the cost of generating training data, we can replace WGBS and TAB-seq with methylation arrays coupled with bisulfite and TAB-treated samples respectively[115]. It is also feasible to supply other enrichment and restriction enzyme sequencing methods as replacement of DeepH&M inputs, such as replacing hmC-Seal with hMeDIP-seq. However, users need to retrain DeepH&M model when using new input methods.

Using 7-week-old mouse cerebellum data for training DeepH&M model, we demonstrated that the estimated 5hmC and 5mC levels were in high concordance with those estimated by combining TAB-seq and WGBS data. DeepH&M estimated 5hmC levels at 85% concordance with TAB-seq data within 0.1 difference and DeepH&M estimated total methylation level at 91% concordance with WGBS data within 0.25 difference. Furthermore, DeepH&M can be generalizable to other tissues and biological time points. DeepH&M model trained on 7-week-old mouse cerebellum data was able to estimate 5hmC and 5mC levels with high performance for 79-week-old mouse cerebellum (concordance for 5hmC and total methylation is 84% and 92%). Of note, differentially hydroxymethylated regions and differentially methylated regions between 7-week-old and 79-week-old mouse cerebellum can be recapitulated using the estimated 5hmC and 5mC values from DeepH&M for the two ages. However, we report relatively lower performance for 7-week-old mouse cortex (concordance for 5hmC and total methylation is 72% and 92%). The relatively lower performance for cortex may be explained by the rather big global differences of 5hmC distribution in cerebellum and cortex, as the mean

5hmC level is 0.19 in cortex and 0.11 in cerebellum. As one of the caveats of DeepH&M, these data suggest that DeepH&M model cannot be generalized to different tissues when 5hmC levels differ greatly between tissues. Indeed, when we applied our DeepH&M model to mouse fetal cortex (mean 5hmC level of 0.05), the concordance for 5hmC and total methylation is 61% and 94%. The extremely low concordance for 5hmC indicates that mean level of 5hmC should be taken into account when applying trained models to different biological systems. Because of the dynamic range of absolute 5hmC levels in different tissues, the relationships between MeDIP-seq, MRE-seq and hmC-Seal data and 5hmC are different in different tissues, and thus a single DeepH&M model cannot be generalized to all tissues. One way to address this limitation is to categorize tissues into multiple classes based on their 5hmC levels and train a DeepH&M model for each group. The DeepH&M model trained for each group can then be generalized to tissues that have similar 5hmC levels.

3.5 Methods

3.5.1 DeepHM model

DeepH&M model is derived from DeepCpG model, which predicts single-cell DNA methylation states using deep learning[131]. DeepH&M model is composed of 3 modules: a regular neural network-based CpG module, a convolutional neural network-based DNA module and a regular neural network-based joint module (**Fig. 2**). CpG module extracts information from inputs of genomic features and methylation features of a CpG with regular neural network. DNA module takes DNA sequence around a CpG as input and uses convolutional neural network to extract information from DNA sequence. The joint module combines outputs from CpG module and DNA module and predicts 5hmC and 5mC simultaneously with regular neural network.

Unlike CpG module in DeepCpG which is a recurrent neural network, the CpG module in DeepH&M is a regular neural network using two fully connected layers with 100 neurons and ReLU activation function. The inputs for CpG module are genomic features and methylation features (**Supplementary Table 1**) for each CpG. Genomic features include GC percent, CpG density and distance to nearest CpG island. Methylation features include MeDIP-seq, MRE-seq and hmC-Seal signal. Because CpGs in proximity tend to have similar 5hmC and 5mC levels, we also include average signal for the above features in neighboring windows (0-50bp,50-250bp,250-500bp,500-1000bp) around the target CpG.

The structure of our DNA module is the same as DNA module of DeepCpG model except that the activation function in our DNA module is tanh function instead of ReLU function (with two connected layers: layer 1 with 120 neurons and layer 2 with 240 neurons).

Joint module uses two fully connected layers with 100 neurons and ReLU activation function to predict 5hmC and 5mC simultaneously, unlike the joint module in DeepCpG which only predicts DNA methylation.

We used data that has at least 25x coverage from TAB-seq data and 20x coverage from WGBS data for training and validation. The feature data is normalized by Z-score normalization. Because the number of high 5hmC level CpGs was much smaller than those with low hmC levels, we balanced the training set through subsampling and oversampling. We divided CpGs into 9 windows based on 5hmC levels: 0-0.05, 0.05-0.1, 0.1-0.15, 0.15-0.2, 0.2-0.25, 0.25-0.3, 0.3-0.35, 0.35-0.4, 0.4-1 and subsampled CpGs if number of CpGs in the window was higher than a threshold and oversampled CpGs if number of CpGs in the window was less than a threshold. The threshold was chosen as the median of the number of CpGs in 9 windows. Data were randomly split into training set (2 million CpGs), validation set (0.5 million CpGs) and test set (the rest). Model parameters were learnt on the training set by minimizing the L2 loss function. We selected the model that had the smallest loss in the validation set and used the model to predict 5hmC and 5mC for all CpGs.

3.5.2 Tissue sample dissection and genomic DNA extraction.

All procedures were approved by the Washington University Institutional Animal Care and Use Committee. Two male 6-week-old C57BL/6J mice and two male 78-week-old C57BL/6J mice were purchased (The Jackson Laboratory, 000664) and allowed to acclimate in the mouse facility for a week. Cerebellum were dissected following protocol described previously[132] from mice in both age groups while the frontal cortex (from bregma +1.0mm to the base of the olfactory bulb) was dissected as described previously[123] from 7-week-old mice. All tissues were snap frozen in liquid nitrogen immediately after dissection.

Each tissue was cut into two pieces with a sterile razor blade for subsequent DNA and RNA extraction immediately following. For genomic DNA extraction, we followed previously established protocol[133]. In brief, each tissue piece was incubated in 600ul of lysis buffer (50mM Tris-HCl pH 8, 1mM EDTA pH 8, 0.5% SDS, 1mg/ml proteinase K) at 55°C for 4 hours. DNA was purified by phenol/chloroform/isoamyl alcohol extraction followed by ethanol extraction. DNA used for MeDIP-seq was sheared into 100-500bp fragment size with Bioruptor Pico Sonication system while DNA for WGBS and TAB-seq was sheared into 200-600bp fragment size with Covaris E220 -Ultrasonicator.

3.5.3 MeDIP-seq, MRE-seq, and hmC-Seal library construction and data processing

MeDIP-seq libraries were generated as previously described[133] with few modifications. 100ng of sheared DNA was ligated with Illumina adapters and methylation-enriched adapter-ligated DNA fragments were immunoprecipitated with 0.1ug of anti-methylcytidine antibody (Eurogentec, BI-MECY-0100). MeDIP DNA fragments were amplified with Illumina barcodes with NEB 2x PCR master mix (NEB, M0541). MeDIP-seq libraries were sequenced on Illumina NovaSeq 6000 platform.

MRE-seq libraries were generated as previously described[133] with few modifications. In brief, 50ng of genomic DNA was digested by four restriction enzymes (HpaII, HinP1I, AciI, HpyCH4IV) that generate a CG overhang. Adapter ligation was performed with custom Illumina adapters (5'- ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3' and 5'-P-CGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'). Adapter-ligated DNA fragments were amplified with Illumina barcodes with NEB 2x PCR master mix (NEB, M0541) and sequenced on Illumina NovaSeq 6000 platform.

To identify 5hmC-enriched regions, we performed Nano-hmC-Seal[48] on tissue samples. In brief, 50ng genomic DNA was used in the tagmentation reaction. The tagmented DNA was glucosylated by incubating in a 50 µl solution containing 1x glucosylation buffer, 200 µM UDP-Azide-Glucose (Active Motif, 55020), and 5 U T4 ß-glucosyltransferase (Thermofisher, EO0831), at 37°C for 1 hr. After glucosylation, the DBCO-PEG4-Biotin reaction and streptavidin C1 beads pull-down were same as the Nano-hmC-Seal[48]. The beads were washed ten times with 1x binding-washing buffer and twice with ddH2O and were re-suspended in 15ul ddH2O. The captured DNA fragments were amplified and barcoded by PCR using the NEBNext 2x PCR master mix (NEB, M0541). hmC-Seal libraries were sequenced on Illumina NovaSeq 6000 platform.

The reads for MeDIP-seq, MRE-seq, hmC-Seal were aligned to the mm9 reference genome with bwa[134] and then processed by methylQA[133]. The signal for MeDIP-seq, MRE-seq and hmC-Seal at each CpG was the number of reads aligned to that location divided by total reads (million). The average signal for MeDIP-seq, MRE-seq and hmC-Seal in each window was the mean of signal at all bases in that window.

3.5.4 WGBS and TAB-seq library construction and data processing

WGBS and TAB-seq libraries were constructed using 5hmC TAB-Seq Kit (WiseGene, K001) following manufacturer's protocol with few modifications detailed below. 5ug of sheared gDNA were treated with β-glucosyltransferase-based reaction to glucosylate 5hmCs. 400ng of glucosylated DNA was incubated in Tet-based oxidation reaction at 37°C for 1.5 hours. 500ng of glucosylated DNA and 250ng of Tet-oxidized DNA were bisulfite converted using EZ DNA Methylation-Gold Kit (Zymo, D5005) for subsequent WGBS and TAB-seq library construction respectively with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences, 30024). WGBS and TAB-seq libraries were sequenced on Illumina NovaSeq 6000 platform.

The reads for TAB-seq and WGBS data were aligned to mm9 reference genome and processed using Bismark[135]. A statistical method MLML was used to integrate TAB-seq and WGBS data to get consistent 5hmC and 5mC and total methylation[113].

3.5.5 DHMRs and DMRs identification

DHMRs between hmC-Seal datasets were defined by DiffBind[121] with a q-value of 0.01.

DHMRs between TAB-seq datasets and DMRs between WGBS datasets were defined by DSS[136]. Two replicates and smoothing options were used for DSS. The called DHMRs and DMRs were then filtered by requiring a minimal coverage of 10 by TAB-seq and WGBS data and the absolute difference of gold standard 5hmC (for DHMRs) and total methylation (for DMRs) in two datasets over 0.15.

3.5.6 Coverage required to call 5% 5hmC

Based on binomial test with a probability of 2.22% for 5mC nonconversion rate, the p-value for using a coverage of 120 to call 5% 5hmC was calculated in R by binom.test(round(120*0.05), 120, p= 0.0222, alternative= "greater"). The resulted p-value for the test was 0.05184. Therefore, a coverage of 120 was required to called 5% 5hmC at 95% confidence level.

3.5.7 Enrichment of 5hmC in genomic features

Enrichment fold = (#CpG for class A CpGs overlapping genomic feature B / #CpG in class A CpGs) / (#CpG for all classes of CpGs overlapping genomic feature B / #CpG in all classes of CpGs).

3.5.8 mRNA-seq library construction and data processing

Total RNA from tissue samples were extracted using TRIzol Reagent as previously detailed[137]. 500ng of total RNA was processed with Universal Plus mRNA-Seq kit (Nugen, 0508-08) to generate mRNA-seq libraries, which were sequenced on Illumina NovaSeq 6000 platform. mRNA reads were aligned to mm9 reference genome using STAR[138]. Read counts for each gene were obtained using HTSeq[139].

3.5.9 Software availability

DeepH&M tool is available in https://epigenome.wustl.edu/DeepHM/.

3.6 Acknowledgements and Funding

We thank all members in Ting Wang's lab for suggestions on the manuscript and the tool. This work was supported by the National Institutes of Health [R01HG007354, R01HG007175, R01ES024992, U01CA200060, U24ES026699, U01HG009391]; and the American Cancer Society Research Scholar Grant [RSG-14-049-01-DMC]; and H.S.J. was supported by NIGMS [T32 GM007067].

3.7 Author Contributions

Y.H., H.S.J., and T.W. conceptualized and designed the study. Y.H. designed DeepH&M algorithm and performed all computational analysis. H.S.J. and X.X. conducted experiments. M.V. and J.D. dissected mouse tissues. D.L. made the DeepH&M website. Y.H., H.S.J., and T.W. wrote and revised the manuscript with input from all authors.

3.8 Figures & Tables

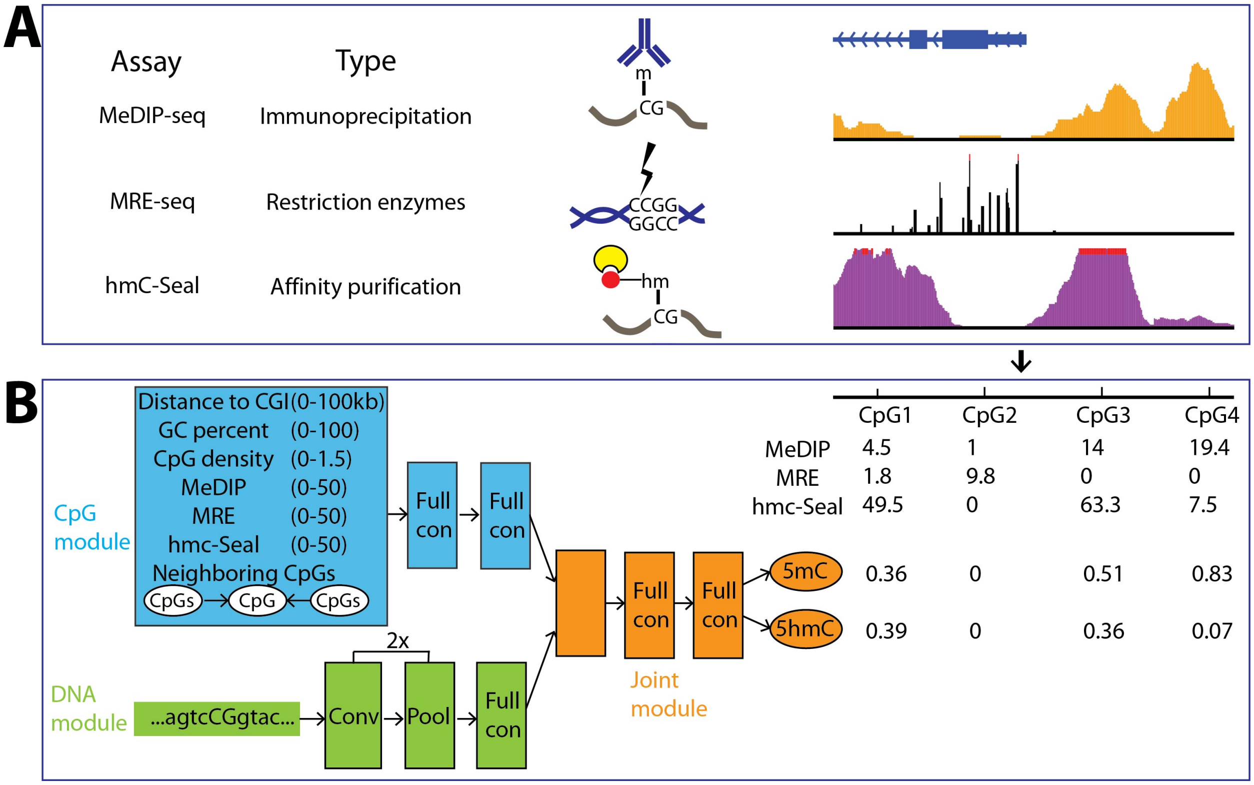


Figure 1. DeepH&M model.

(**A**). Schematic explanations for the 3 main assays used for DeepH&M model. (**B**). Structure of DeepH&M model. DeepH&M is composed of 3 modules. CpG module takes inputs of genomic features and methylation features. DNA module processes raw DNA sequence data using a convolutional neural network. Joint module combines outputs from CpG module and DNA module to predict 5hmC and 5mC simultaneously. Examples were given to show how 5hmC and 5mC were predicted from the 3 main assays. Conv is convolutional layer. Pool is pooling layer. Full con is full connected layer.

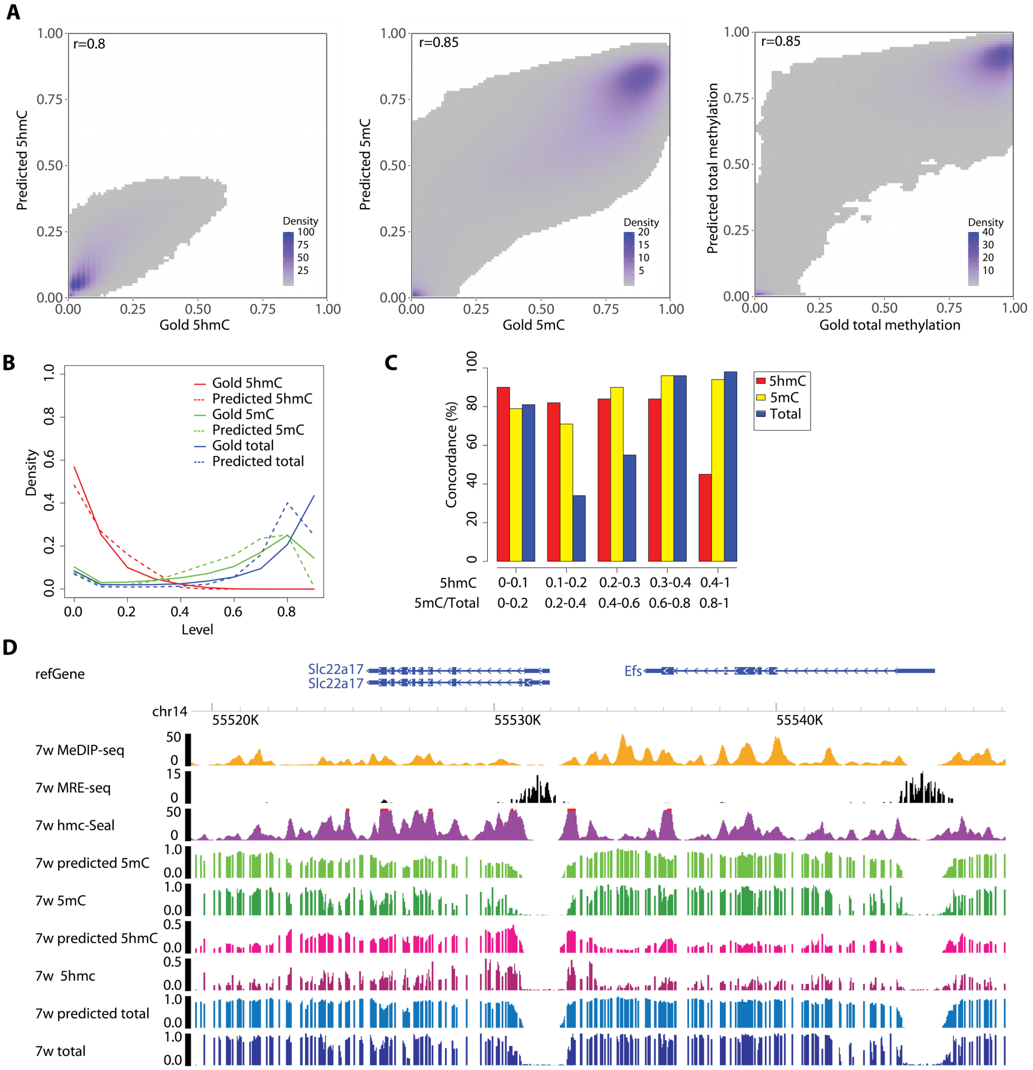


Figure 2. Performance of DeepH&M model in 7-week-old mouse cerebellum.

(**A**). Density plots of predictions and gold standard data for 5hmC, 5mC and total methylation. Pearson correlation coefficient is used as correlation metric. (**B**). Global distribution comparison of predictions and gold standard data for 5hmC, 5mC and total methylation. (**C**). Concordance between predictions and gold standard data for 5hmC, 5mC and total methylation at CpGs with differing 5hmC/5mC/total methylation levels. For 5hmC, 0.1 difference is used to calculate concordance. For 5mC and total methylation, 0.25 difference is used. Concordance for five ascending 5hmC windows and five ascending 5mC/total methylation windows are calculated to see how concordance distributes in differing 5hmC/5mC/total methylation levels. (**D**). Genome browser view of predictions and gold standard data for 7-week-old cerebellum at a representative locus.

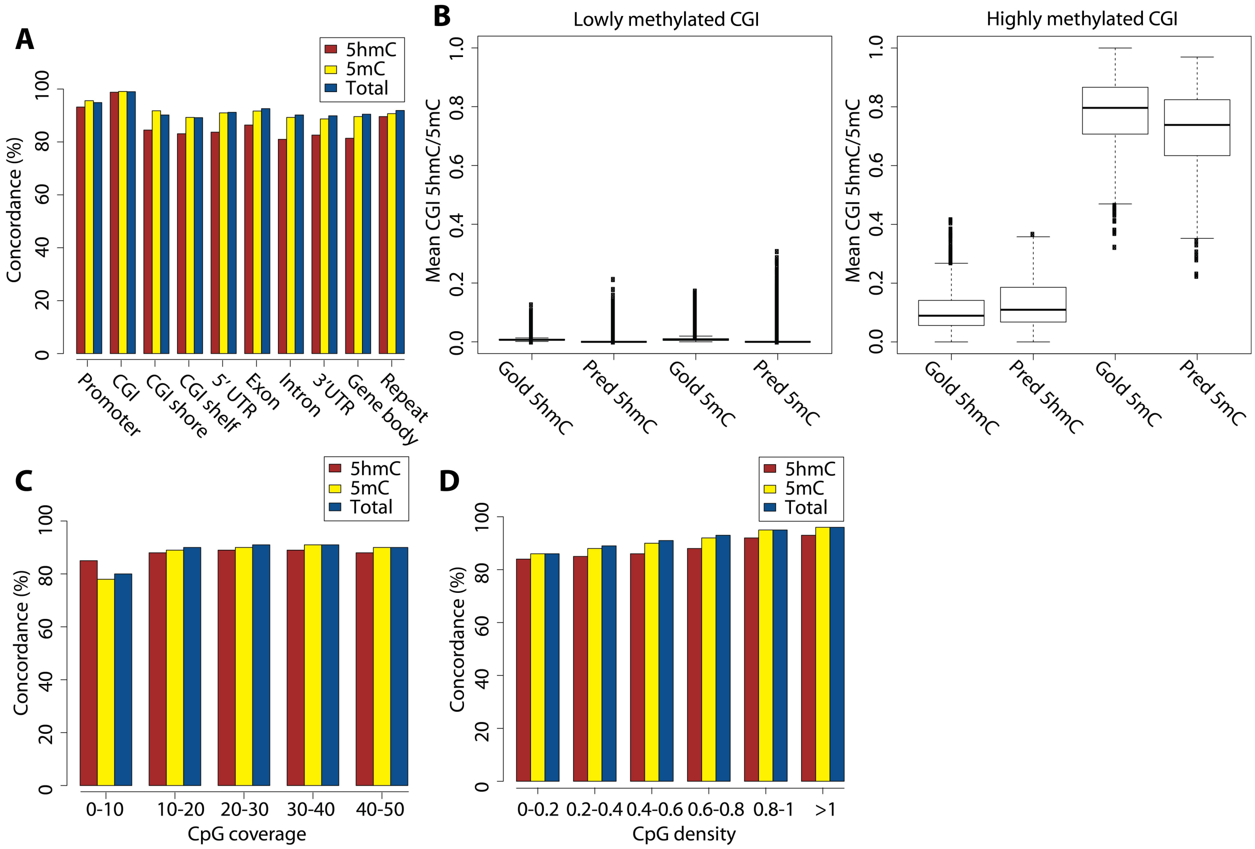


Figure 3. Factors affecting concordance between gold standard data and predictions.

(**A**). Concordance for 5hmC/5mC/total methylation at different genomic features. (**B**). Comparison of gold standard 5hmC/5mC and predicted 5hmC/5mC at lowly methylated CGIs and highly methylated CGIs. CGIs are divided into lowly methylated CGIs (<0.2) and highly methylated CGIs (>0.7) based on their average total methylation levels. (**C**). Concordance for 5hmC/5mC/total methylation as a function of CpG coverage. For 5hmC concordance, CpG coverage is from TAB-seq data. For 5mC/total methylation concordance, CpG coverage is from WGBS data. (**D**). Concordance for 5hmC/5mC/total methylation as a function of CpG density.

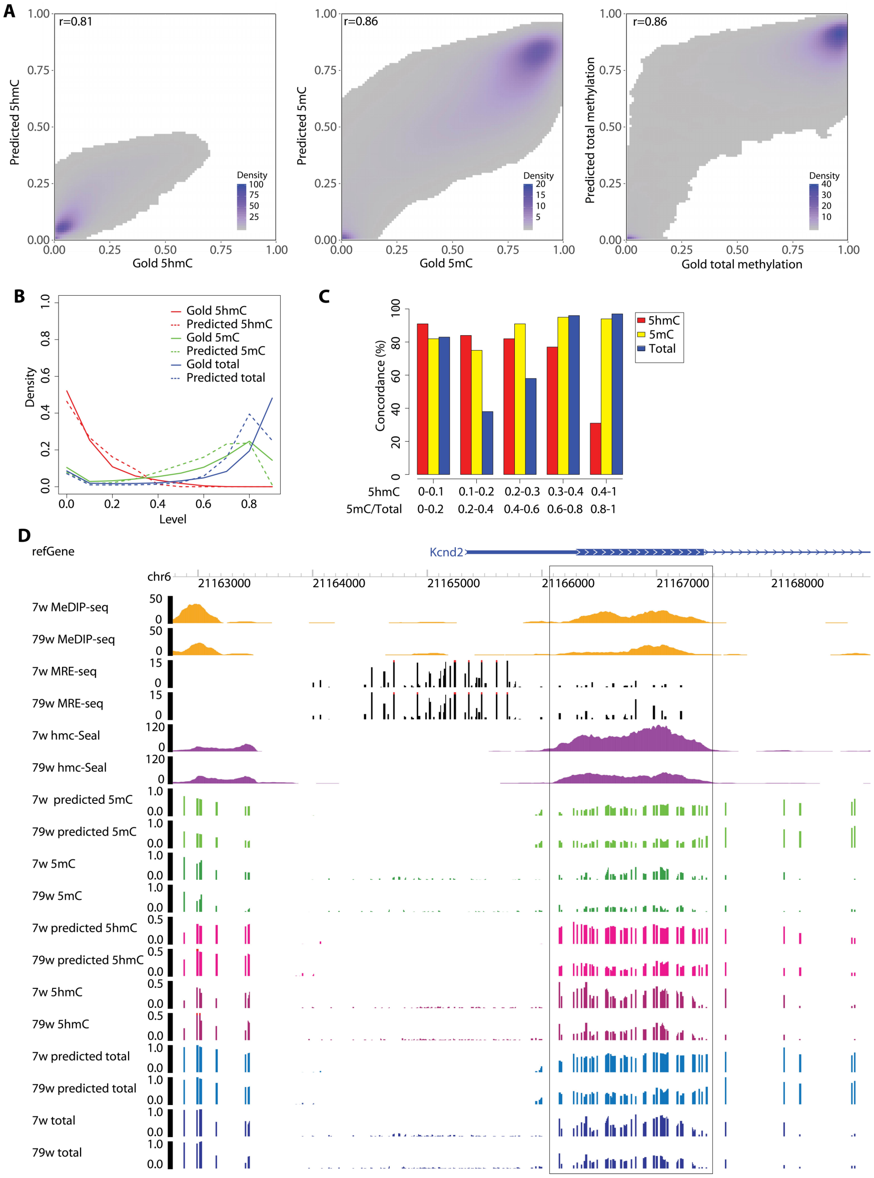


Figure 4. Performance of DeepH&M model in 79-week-old mouse cerebellum.

(**A**). Density plots of predictions and gold standard data for 5hmC, 5mC and total methylation. (**B**). Global distribution comparison of predictions and gold standard data for 5hmC, 5mC and total methylation. (**C**). Concordance between predictions and gold standard data for 5hmC, 5mC and total methylation at CpGs with differing 5hmC/5mC/total methylation levels. (**D**). Genome browser view of a differentially hydroxymethylated region between 7-week-old and 79-week-old cerebellum. The selected box is the DHMR. The 5hmC changes at this region is supported by changes of gold standard 5hmC, predicted 5hmC and also hmC-Seal signal between the two ages.

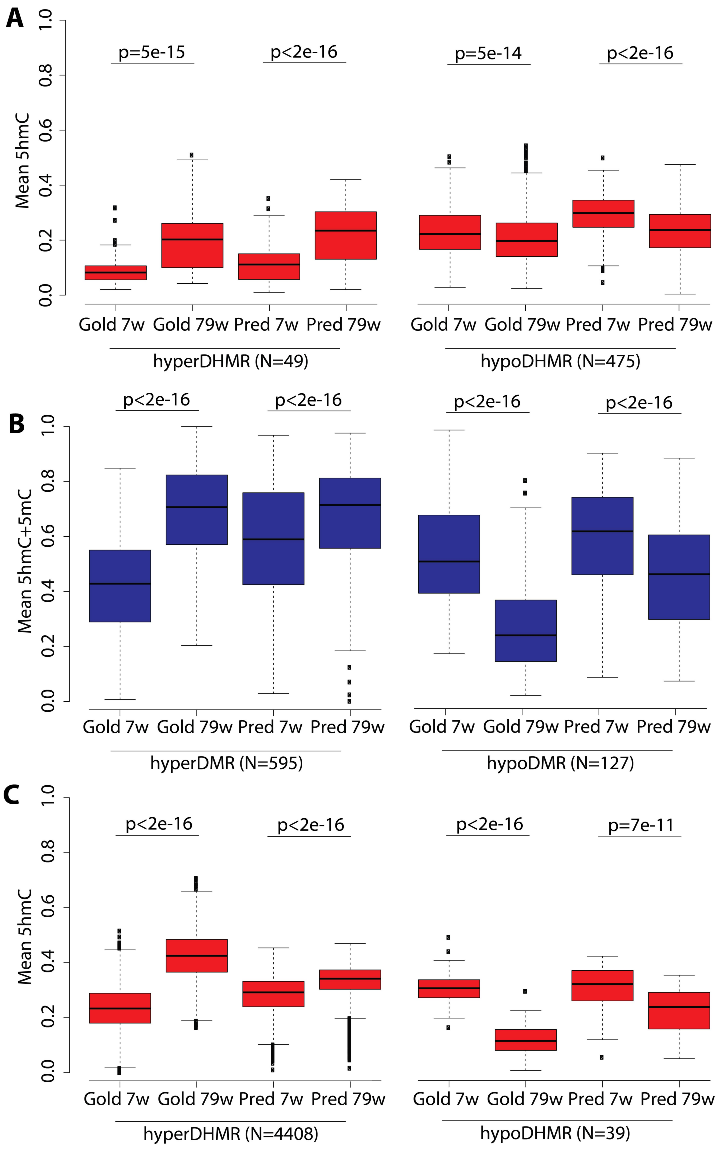


Figure 5. DeepH&M can predict differentially hydroxymethylated regions and differentially methylated regions between 7-week-old and 79-week-old mouse cerebellum.

(**A**). Distribution of mean 5hmC for gold standard data and predictions at hyperDHMRs and hypoDHMRs defined by hmC-Seal data between 7w and 79w cerebellum. gold is for gold standard data. pred is for prediction. N is the number. (**B**). Distribution of mean 5hmC+5mC for gold standard data and predictions at hyperDMRs and hypoDMRs defined by WGBS data between 7w and 79w cerebellum. (**C**). Distribution of mean 5hmC for gold standard data and predictions at hyperDMRs and hypoDMRs defined by TAB-seq data between 7w and 79w cerebellum.

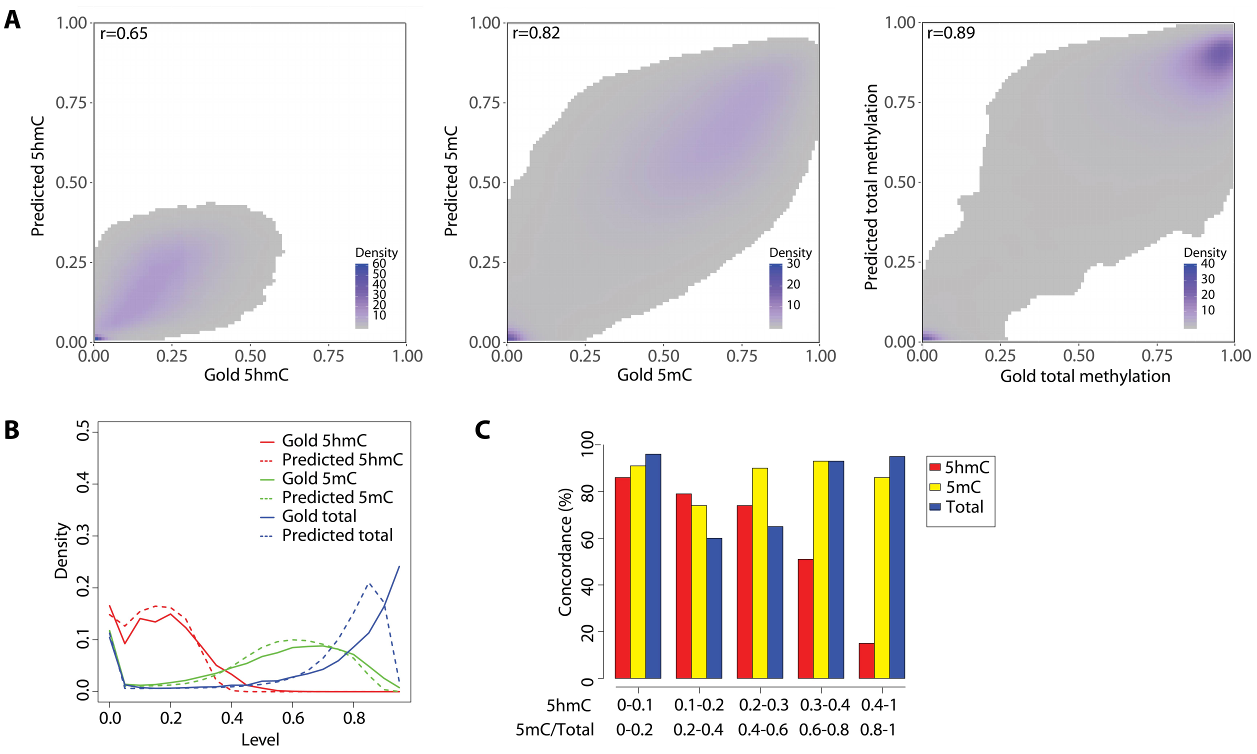
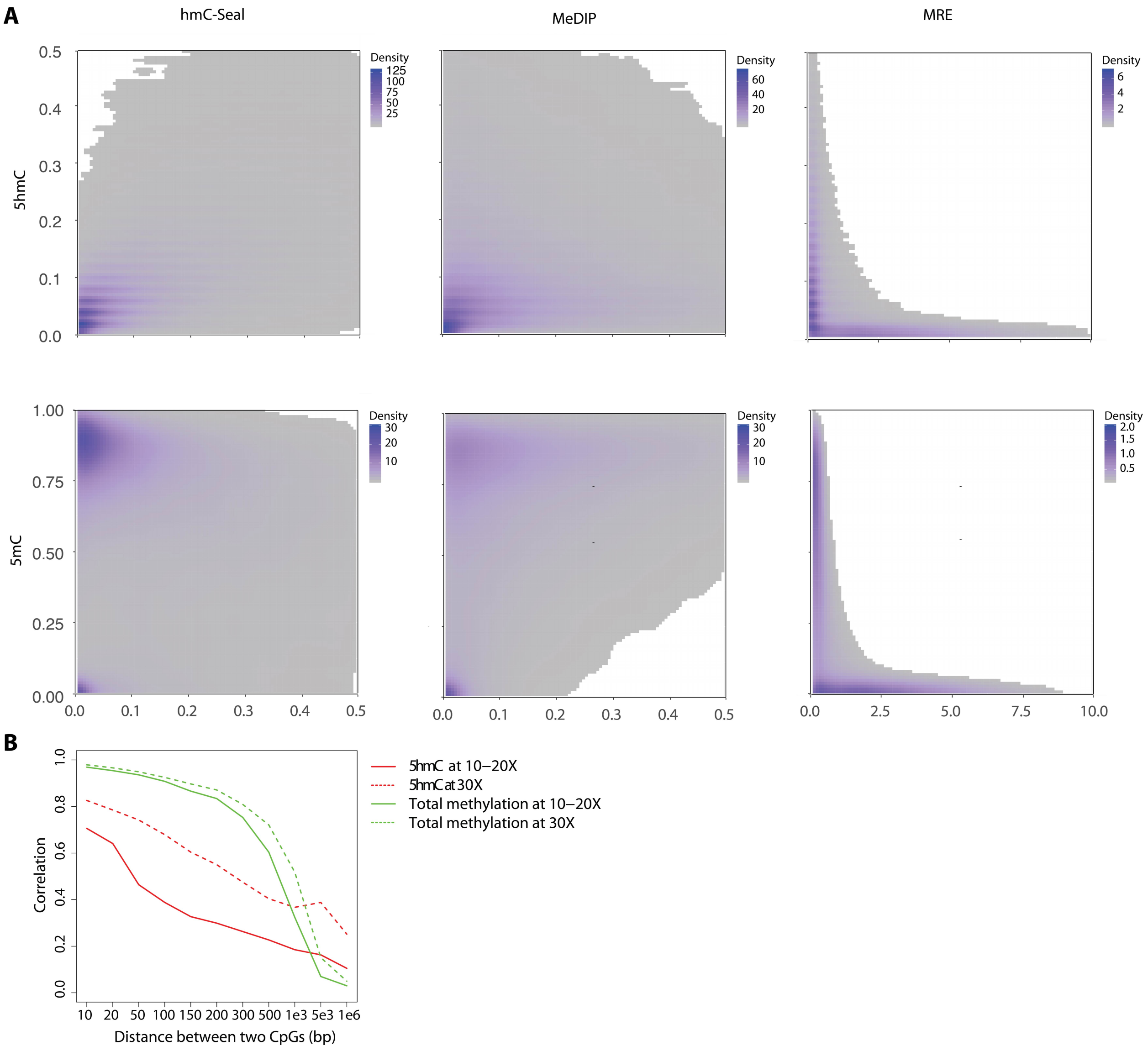


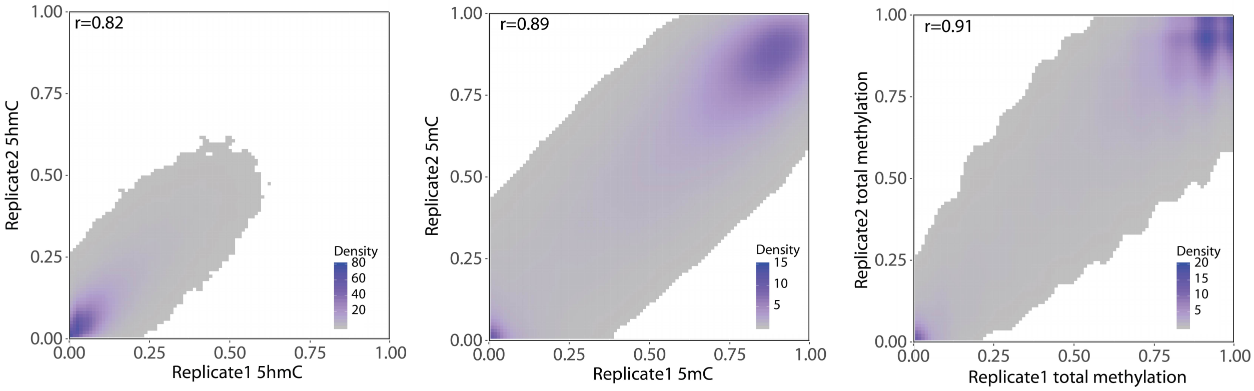
Figure 6. Performance of DeepH&M model in 7-week-old mouse cortex.

(**A**). Density plots of predictions and gold standard data for 5hmC, 5mC and total methylation. (**B**). Global distribution comparison of predictions and gold standard data for 5hmC, 5mC and total methylation. (**C**). Concordance between predictions and gold standard data for 5hmC, 5mC and total methylation at CpGs with differing 5hmC/5mC/total methylation levels.



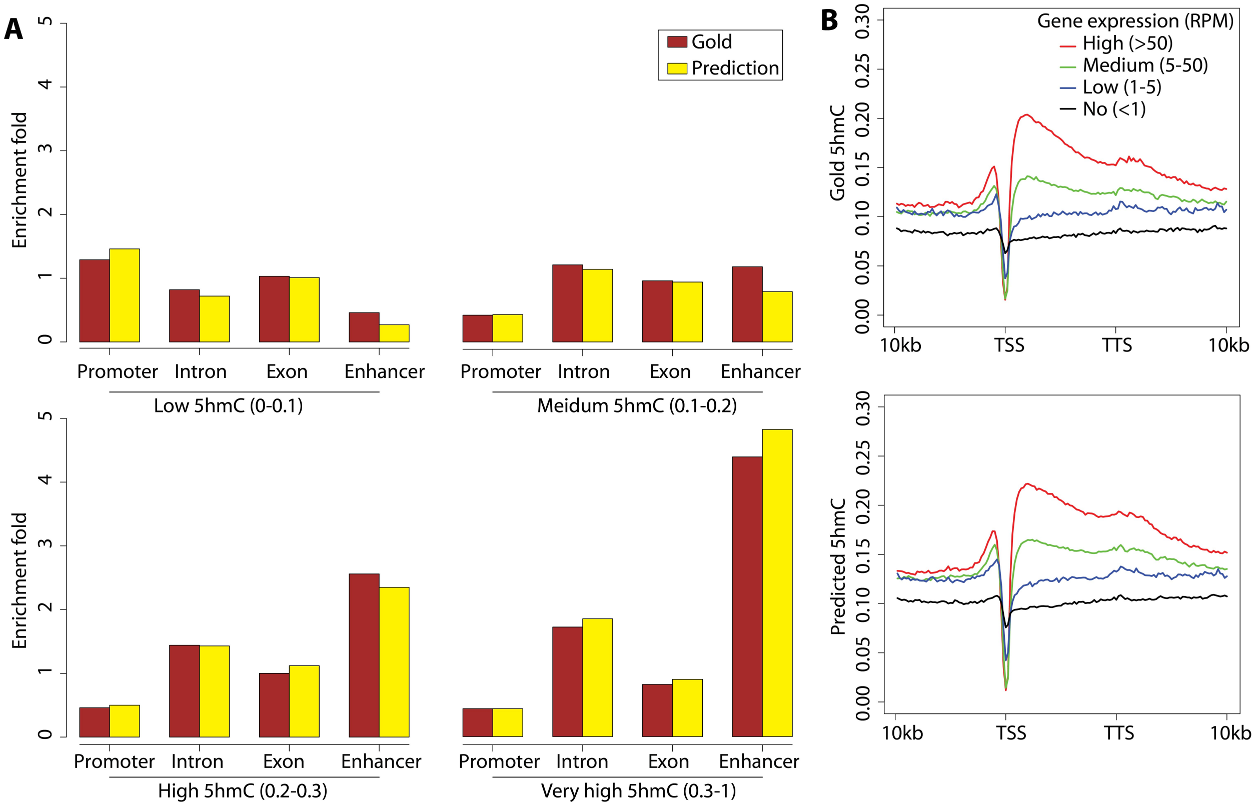
Supplementary Figure 1. Relationship between main DeepH&M features and gold standard 5hmC and 5mC.

(**A**). Density plots of 5hmC/5mC as a function of hmC-Seal, MeDIP-seq, MRE-seq signal at the CpG sites. X-axis represents feature signals. Y-axis represents gold standard 5hmC/5mC values. Color bar shows the density of points. (**B**). Correlation of 5hmC and total methylation levels between two CpG sites as a function of distance between the two CpG sites. Two coverage cutoffs (10-20X, >30X) are chosen to demonstrate how the correlation between two CpG sites can be dependent on sequencing coverage (which is a surrogate of confidence in predicted methylation levels). Pearson correlation coefficient is used as correlation metric.



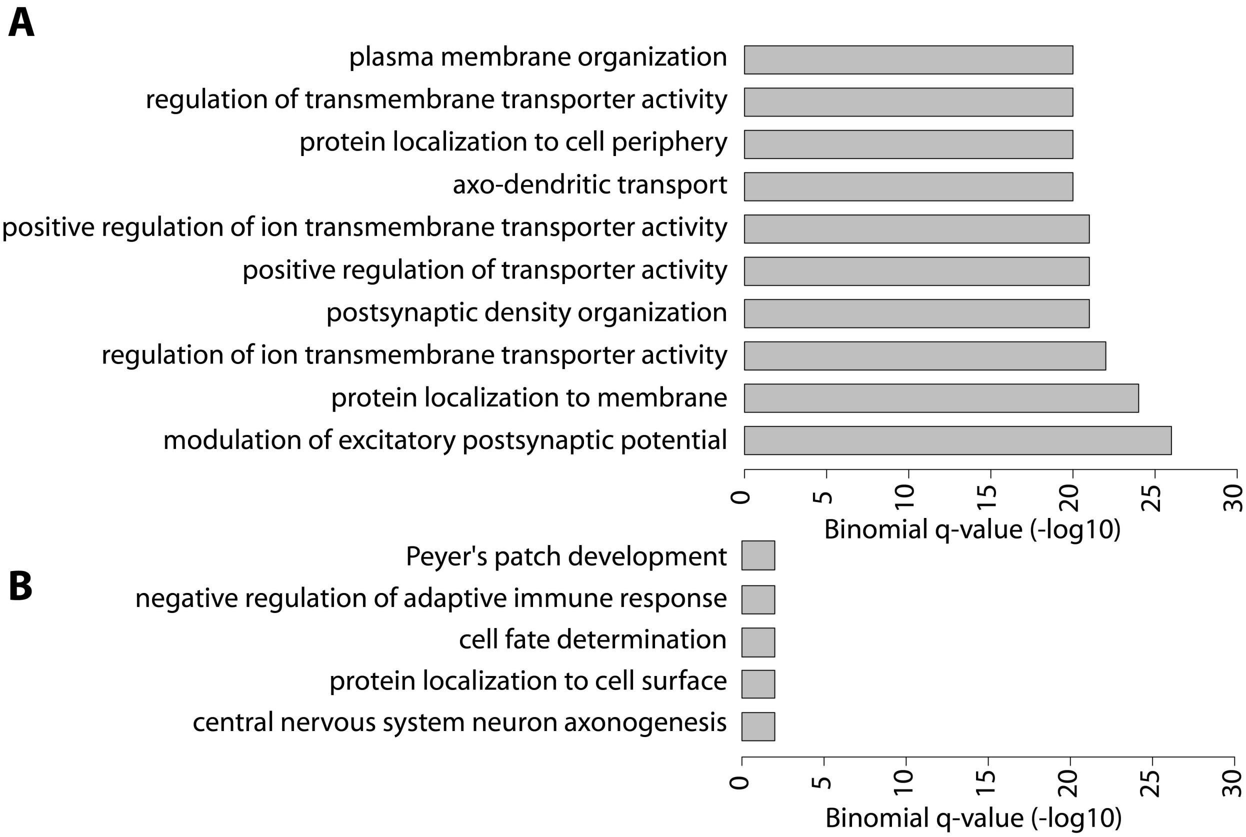
Supplementary Figure 2. Density plots of two 7-week-old cerebellum replicates data for 5hmC, 5mC and total methylation.

Pearson correlation coefficient is used as correlation metric.



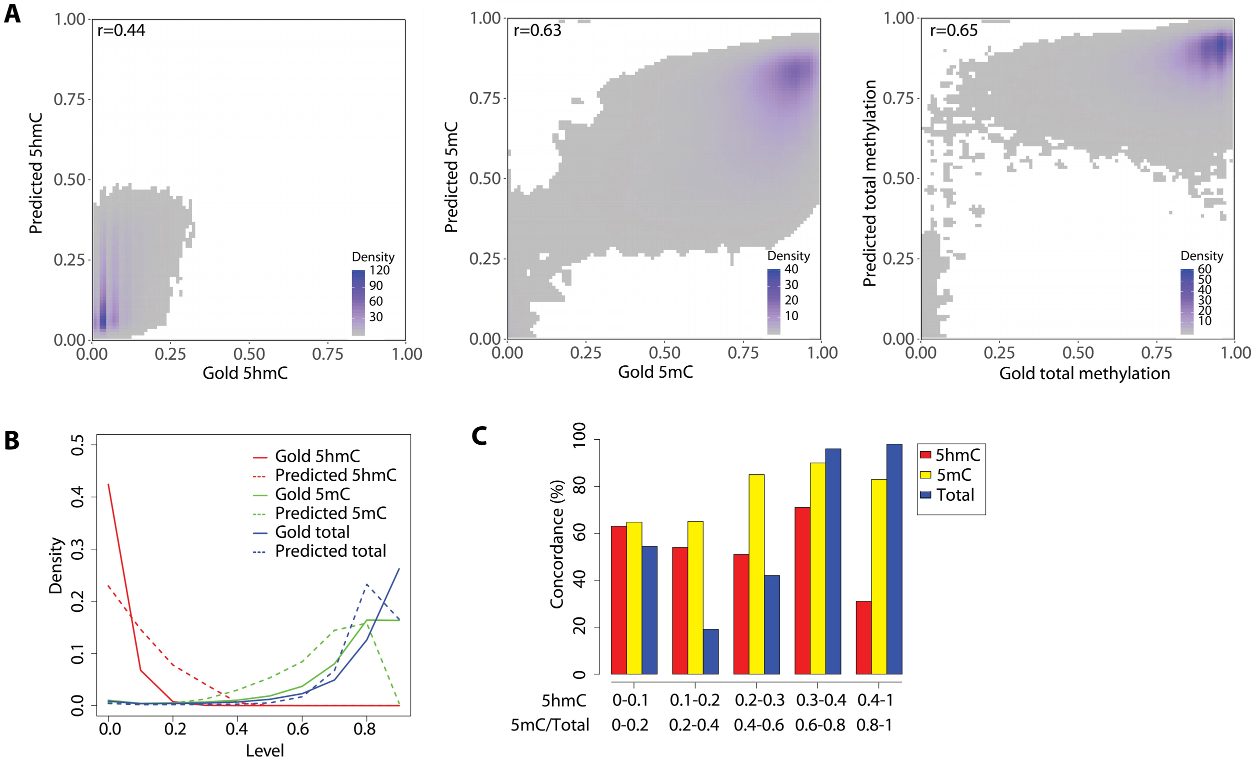
Supplementary Figure 3. Comparison between gold standard 5hmC and 5hmC predictions for known 5hmC function.

(**A**). Enrichment of gold standard 5hmC and 5hmC predictions at genomic features. CpGs are divided into four groups based on their 5hmC levels and enrichment is calculated for each group. (**B**). Average 5hmC profiles for gold standard 5hmC and 5hmC predictions across gene body of genes with differing expression levels. Genes are divided into four groups based on their expression levels at 7-week-old cerebellum. RPM (reads per million) is used for gene expression normalization. TSS: transcription start site. TTS: transcription termination site.



Supplementary Figure 4. Enrichment of biological processes for hyperDHMRs and hyperDMRs between 7-week-old and 79-week-old cerebellum using GREAT.

(**A**). Enrichment of biological processes for hyperDHMRs. Default settings were used for GREAT. (**B**). Enrichment of biological processes for hyperDMRs.



Supplementary Figure 5. Performance of DeepH&M model in fetal mouse cortex.

(**A**). Density plots of predictions and gold standard data for 5hmC, 5mC and total methylation. (**B**). Global distribution comparison of predictions and gold standard data for 5hmC, 5mC and total methylation. (**C**). Concordance between predictions and gold standard data for 5hmC, 5mC and total methylation at CpGs with differing 5hmC/5mC/total methylation levels.

Feature	Description	Motivation
distance to nearest CGI	help distinguish CGI, CGI shore and non-CGI	CGIs tend to be significantly unmethylated and CGI shores tend to be variably methylated
GC percent	GC percent in multiple upstream and downstream windows (upstream and downstream 0-50bp,50-250bp,250-500bp,500-1000bp)	higher GC content empirically shows lower methylation
CpG density	CpG density in multiple upstream and downstream windows	higher GC content empirically shows lower methylation
MeDIP signal	MeDIP signal in multiple upstream and downstream windows	MeDIP-seq measures the enrichment of methylation
MRE signal	MRE signal in multiple upstream and downstream windows	MRE-seq measures the enrichment of unmethylation at enzyme cut sites
MRE site	whether CpG is in restriction enzyme cut sites	MRE-seq measures the enrichment of unmethylation at enzyme cut sites
hmC-Seal signal	hmC-Seal signal in multiple upstream and downstream windows	hmC-Seal measures the enrichment of hydroxymethylation
DNA sequence	DNA sequence in upstream and downstream 500bp	Specific motifs in the sequence can be subjected to methylation and hydroxymethylation

Supplementary Table 1. Features for DeepH&M model.

	7w replicate1	7w replicate2	79w replicate1	79w replicate2
mean 5hmC	0.107	0.114	0.132	0.133
mean 5mC	0.639	0.64	0.637	0.634

Supplementary Table 2. Mean 5hmC and 5mC levels in 7w and 79w mouse cerebellum.

About 4 million CpGs are considered for the analysis by requiring 10x coverage for WGBS and 25x coverage for TAB-seq in all cerebellum samples.

Chapter 4: Conclusions and Future Directions

Yu He

4.1 Significance

In this dissertation, I present an online tool EpiCompare, which compares different epigenomes in order to identify regions with epigenomic features specific to certain types of tissues or cells, and a deep learning-based algorithm DeepH&M, which integrates enrichment and restriction enzyme sequencing methods to estimate the absolute levels of hydroxymethylation and methylation at single CpG resolution.

We have showed that the EpiCompare can easily identify regulatory elements such as enhancers, promoters, and regions occupied by epigenetic features that are unique to a specific tissue or cell type, as well as those that are shared by multiple tissue and cell types. Our tool is designed specifically for biologists in such a way that no programming or data processing capacity is required to perform genome-wide analysis. We demonstrated that our tool could identify endoderm-specific enhancers and analysis on these enhancers revealed the regulatory network common to all endoderm tissues. In identifying regions with epigenomic features specific to combinations of tissue or cell types, EpiCompare has several advantages over existing methodologies reported in the FANTOM5, Roadmap, and others. First, investigators can compare enhancers, promoters, and specific histone marks using any combination of tissue and cell types depending on their needs. This enables the identification of specific epigenomic features associated with specific biological entities, such as lineage-specific enhancers. Second, the tool is user-friendly so that an experimental biologist with little or no programming experience can easily use. Investigators can test a variety of hypotheses by designing specific combinations of epigenome comparisons using Roadmap data and/or their own data, and EpiCompare provides a quality assessment of the predictions. The predicted regions can be readily visualized and further explored using the WashU Epigenome Browser.

We have showed that using 7-week-old mouse cerebellum data for training DeepH&M model, the 5hmC and 5mC levels predicted by DeepH&M were in high concordance with whole genome bisulfite- based approaches. The DeepH&M model can be applied to 7-week old frontal cortex and 79-week cerebellum revealing the robust generalizability of this method to other tissues from various biological time points. Currently, TAB-seq and oxidative-bisulfite sequencing (oxBS-seq) are the gold standard methods for providing single-CpG-resolution DNA hydroxymethylomes[46,125]. These two methods require very high coverage to confidently call 5hmC at all cytosines. The coverage required for oxBS-seq is even higher due to the fact that oxBS-seq measures 5hmC indirectly through subtracting measured 5mC from measured total methylation. The high cost associated with the high coverage is a significant barrier for individual laboratories to adopt TAB-seq and oxBS-seq as a routine assay for DNA hydroxymethylomes. Indeed, so far only a few cell types have deeply sequenced hydroxymethylomes at single-base resolution[46,115,123,124,126–130]. Our algorithm DeepH&M can overcome this potential cost-barrier problem. The cost of the three assays combined is <5% of WGBS and TAB-seq. About 50-100 million MeDIP reads, 30 million MRE reads and 50 million hmC-Seal-seq reads are sufficient for measuring a hydroxymethylome with DeepH&M, which translates to roughly 3x coverage of the human or mouse genome. Also, TAB-seq requires ~3ug of genomic DNA while MeDIP-seq, MRE-seq, and hmC-Seal can be generated from 100ng or less input thus allowing DeepH&M to be more amenable to rare or difficult-to-procure cells or samples. Compared to 100x coverage for TAB-seq and 20x coverage for WGBS, our method can minimize the cost of generating a complete hydroxymethylome by 40-fold. Furthermore, DeepH&M can estimate for all CpGs while WGBS and TAB-seq miss a significant fraction of the genome due to low coverage. As mentioned previously, previous TAB-seq study on H1 cells could only confidently call 20% or higher 5hmC at a coverage of 27 and thus identified less than 1 million hydroxymethylated CpGs[46].

4.2 Future Directions

EpiCompare has some limitations. First, the regulatory elements used in this tool are defined based on the ChromHMM model. Although considered the state-of-the-art, ChromHMM model still has limited sensitivity and specificity, especially for identifying enhancers[76]. The performance of predicting tissue or cell type-specific enhancers is clearly dependent on the performance of ChromHMM. Second, EpiCompare is based on comparison of binary data including chromatin states and histone mark peaks. It could potentially miss regions with quantitatively different signal between samples. For example, it could not distinguish a weak enhancer from a strong enhancer if both had signals over the threshold. It could also not distinguish two quantitatively different weak enhancers which were below the thresh-old. These cases are false negatives for EpiCompare. The comparison of binary data can also lead to false positives if two samples had very similar signal at one region, with one above the threshold and the other below the threshold. Third, we implemented three very simple statistical models, and potentially could oversimplify the problem of identifying tissue or cell type-specific features. Frequency cutoff method uses simple cutoffs, and Fisher's exact test assumes the occurrence of features as hypergeometric distribution while k-means clustering method assumes certain number of clusters in the data and groups them based on similarity. All of them assume the independence of samples, but biological samples are clearly not independent from each other. The statistical models also do not consider the distribution of each feature along the genome of each sample. However, we are encouraged by the strong performance of these simple models, and anticipate that development of more sophisticated models in the future will surely improve the accuracy of feature identification.

DeepH&M model trained on 7-week-old mouse cerebellum data was able to estimate 5hmC and 5mC levels with high performance for 79-week-old mouse cerebellum (concordance for 5hmC and total methylation is 84% and 92%). However, we report relatively lower performance for 7-week-old mouse cortex (concordance for 5hmC and total methylation is 72% and 92%). The relatively lower performance for cortex may be explained by the rather big global differences of 5hmC distribution in cerebellum and cortex, as the mean 5hmC level is 0.19 in cortex and 0.11 in cerebellum. As one of the caveats of DeepH&M, these data suggest that DeepH&M model cannot be generalized to different tissues when 5hmC levels differ greatly between tissues. Indeed, when we applied our DeepH&M model to mouse fetal cortex (mean 5hmC level of 0.05), the concordance for 5hmC and total methylation is 61% and 94%. The extremely low concordance for 5hmC indicates that mean level of 5hmC should be taken into account when applying trained models to different biological systems. Because of the dynamic range of absolute 5hmC levels in different tissues, the relationships between MeDIP-seq, MRE-seq and hmC-Seal data and 5hmC are different in different tissues, and thus a single DeepH&M model cannot be generalized to all tissues. One way to address this limitation is to categorize tissues into multiple classes based on their 5hmC levels and train a DeepH&M model for each group. The DeepH&M model trained for each group can then be generalized to tissues that have similar 5hmC levels. In the future we need to collect methylation and hydroxymethylation datasets for different tissues as training sets and build complete models for diverse tissues with mean 5hmC levels at all ranges. So far only a few cell types have deeply sequenced hydroxymethylomes at single-base esolution[46,115,123,124,126–130]. These include ES cells, neurons, cortex, cerebellum, olfactory bulb, lung, liver. However, the corresponding MeDIP-seq, MRE-seq and hmc-Seal data for these cell and tissue types were missing. We can collaborate with other labs in generating these feature data for these cell and tissue types. For other cell and tissue types, the burden may be on individual labs to generate the complete set of data required for training DeepH&M.

References

1. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* (2007) doi:10.1016/j.cell.2007.02.005.
2. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* (2011) doi:10.1016/j.cell.2011.08.008.
3. Tian, Z. *et al.* Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.* (2012) doi:10.1186/gb-2012-13-10-R86.
4. Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* (2007) doi:10.1016/j.cell.2007.01.033.
5. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* (2007) doi:10.1016/j.cell.2007.05.009.
6. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* (2010) doi:10.1073/pnas.1016071107.
7. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* (2007) doi:10.1016/j.cell.2007.05.042.
8. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* (2007) doi:10.1038/ng1966.
9. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* (2007) doi:10.1038/nature06008.
10. Bannister, A. J. *et al.* Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. *J. Biol. Chem.* (2005) doi:10.1074/jbc.M500796200.
11. Bernstein, B. E. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* (2005) doi:10.1016/j.cell.2005.01.001.
12. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* (2011) doi:10.1038/nature09725.
13. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* (2006) doi:10.1016/j.cell.2006.02.041.
14. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* (2000) doi:10.1038/47412.
15. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* (2011) doi:10.1038/nature09906.

16. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* (2010) doi:10.1038/nbt.1662.

17. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* (2012) doi:10.1038/nmeth.1937.

18. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gks1284.

19. Song, J. & Chen, K. C. Spectacle: Fast chromatin state annotation using spectral learning. *Genome Biol.* (2015) doi:10.1186/s13059-015-0598-0.

20. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* (2015) doi:10.1038/nature14248.

21. Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* (2010) doi:10.1016/j.cell.2010.09.009.

22. Lai, W. K. M. & Buck, M. J. An integrative approach to understanding the combinatorial histone code at functional elements. *Bioinformatics* (2013) doi:10.1093/bioinformatics/btt382.

23. Sequeira-Mendes, J. *et al.* The functional topography of the Arabidopsis genome is organized in a reduced number of linear motifs of chromatin states. *Plant Cell* (2014) doi:10.1105/tpc.114.124578.

24. Sohn, K. A. *et al.* HiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv117.

25. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* (2015) doi:10.1038/nature13835.

26. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* (2001) doi:10.1038/35057062.

27. Myers, R. M. *et al.* A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biol.* (2011) doi:10.1371/journal.pbio.1001046.

28. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* (2009) doi:10.1038/nature07829.

29. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* (2012) doi:10.1038/nature11243.

30. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484–492 (2012).

31. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* **11**, 191–203 (2010).

32. Robertson, K. D. DNA methylation and human disease. *Nat Rev Genet* **6**, 597–610 (2005).

33. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465–476 (2008).

34. Smith, Z. D. & Meissner, A. DNA methylation: Roles in mammalian development. *Nature Reviews Genetics* (2013) doi:10.1038/nrg3354.

35. Lowdon, R. F. *et al.* Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. *Nat. Commun.* (2014) doi:10.1038/ncomms6442.

36. Zhang, B. *et al.* Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res.* (2013) doi:10.1101/gr.156539.113.

37. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (80-.).* **324**, 929–930 (2009).

38. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37**, e123 (2009).

39. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (80-.).* **333**, 1300–1303 (2011).

40. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334–35338 (2011).

41. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (80-.).* **333**, 1303–1307 (2011).

42. Song, C. X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* **30**, 1107–1116 (2012).

43. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (80-.).* **324**, 930–935 (2009).

44. Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68–72 (2011).

45. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).

46. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).

47. Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68–72 (2011).

48. Han, D. *et al.* A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Mol Cell* **63**, 711–719 (2016).

49. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817–825 (2010).

50. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).

51. Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* **9**, e1002968 (2013).

52. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473–476 (2012).

53. Sohn, K. A. *et al.* hiHMM: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics* **31**, 2066–2074 (2015).

54. Biesinger, J., Wang, Y. & Xie, X. Discovering and mapping chromatin states using a tree hidden Markov model. *BMC Bioinformatics* **14 Suppl 5**, S4 (2013).

55. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

56. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

57. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

58. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).

59. Won, K. J. *et al.* Comparative annotation of functional regions in the human genome using epigenomic data. *Nucleic Acids Res* **41**, 4423–4432 (2013).

60. Mahony, S. *et al.* An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput Biol* **10**, e1003501 (2014).

61. Ji, H., Li, X., Wang, Q. F. & Ning, Y. Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci U S A* **110**, 6789–6794 (2013).

62. Yen, A. & Kellis, M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat Commun* **6**, 7973 (2015).

63. Bock, C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* **13**, 705–719 (2012).

64. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320–331 (2010).

65. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).

66. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**, 853–862 (2005).

67. Xiao, Y. *et al.* MeSiC: A Model-Based Method for Estimating 5 mC Levels at Single-CpG Resolution from MeDIP-seq. *Sci Rep* **5**, 14699 (2015).

68. Stevens, M. *et al.* Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res* **23**, 1541–1553 (2013).

69. Pavlovic, M. *et al.* DIRECTION: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. *Bioinformatics* **33**, 2986–2994 (2017).

70. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

71. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science (80-.).* **281**, 60–63 (1998).

72. Sakabe, N. J., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *Genome Biol* **13**, 238 (2012).

73. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**, 895–907 (2015).

74. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

75. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827–841 (2013).

76. Song, J. & Chen, K. C. Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol* **16**, 33 (2015).

77. Zhou, X. *et al.* Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat Biotechnol* **33**, 345–346 (2015).

78. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311–318 (2007).

79. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

80. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and

predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931–21936 (2010).

81. Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).

82. Pan, J. B. *et al.* PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. *PLoS One* **8**, e80747 (2013).

83. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

84. Zorn, A. M. & Wells, J. M. Vertebrate endoderm development and organ formation. *Annu Rev Cell Dev Biol* **25**, 221–251 (2009).

85. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–589 (2010).

86. Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279–289 (2002).

87. Gao, N. *et al.* Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev* **22**, 3435–3448 (2008).

88. Gosalia, N., Yang, R., Kerschner, J. L. & Harris, A. FOXA2 regulates a network of genes involved in critical functions of human intestinal epithelial cells. *Physiol Genomics* **47**, 290–297 (2015).

89. Lee, C. S., Friedman, J. R., Fulmer, J. T. & Kaestner, K. H. The initiation of liver development is dependent on Foxa transcription factors. *Nature* **435**, 944–947 (2005).

90. Wan, H. *et al.* Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J Biol Chem* **280**, 13809–13816 (2005).

91. DeLaForest, A. *et al.* HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells. *Development* **138**, 4143–4153 (2011).

92. Pontoglio, M. Hepatocyte nuclear factor 1, a transcription factor at the crossroads of glucose homeostasis. *J Am Soc Nephrol* **11 Suppl 16**, S140-3 (2000).

93. Yang, R., Kerschner, J. L. & Harris, A. Hepatocyte nuclear factor 1 coordinates multiple processes in a model of intestinal epithelial cell function. *Biochim Biophys Acta* **1859**, 591–598 (2016).

94. Spence, J. R. *et al.* Directed differentiation of human pluripotent stem cells into intestinal tissue in vitro. *Nature* **470**, 105–109 (2011).

95. Lee, H. J. *et al.* Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nat Commun* **6**, 6315 (2015).

96. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res* **37**, D623-8 (2009).

97. RStudio Inc. Shiny: Easy web applications in R. *http://shiny.rstudio.com/* (2014).

98. Zhou, X. *et al.* The Human Epigenome Browser at Washington University. *Nat Methods* **8**, 989–990 (2011).

99. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

100. Leisch, F. A toolbox for K-centroids cluster analysis. *Comput. Stat. Data Anal.* (2006) doi:10.1016/j.csda.2005.10.006.

101. Kodinariya, T. M. & Makwana, P. R. Review on determining number of Cluster in K-Means Clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* (2013).

102. Chang, C. W. *et al.* Identification of human housekeeping genes and Tissue-Selective genes by microarray Meta-Analysis. *PLoS One* (2011) doi:10.1371/journal.pone.0022859.

103. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* (2005) doi:10.1093/bioinformatics/bti042.

104. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155**, 39–55 (2013).

105. Chen, T. & Dent, S. Y. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet* **15**, 93–106 (2014).

106. Pfeifer, G. P., Xiong, W., Hahn, M. A. & Jin, S. G. The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res* **356**, 631–641 (2014).

107. Greco, C. M. *et al.* DNA hydroxymethylation controls cardiomyocyte gene expression in development and hypertrophy. *Nat Commun* **7**, 12418 (2016).

108. Jeschke, J., Collignon, E. & Fuks, F. Portraits of TET-mediated DNA hydroxymethylation in cancer. *Curr Opin Genet Dev* **36**, 16–26 (2016).

109. Smeets, E. *et al.* The role of TET-mediated DNA hydroxymethylation in prostate cancer. *Mol Cell Endocrinol* **462**, 41–55 (2018).

110. Monticelli, S. DNA (Hydroxy)Methylation in T Helper Lymphocytes. *Trends Biochem Sci* **44**, 589–598 (2019).

111. Wu, H. & Zhang, Y. Charting oxidized methylcytosines at base resolution. *Nat Struct Mol Biol* **22**, 656–661 (2015).

112. Szyf, M. The elusive role of 5’-hydroxymethylcytosine. *Epigenomics* **8**, 1539–1551 (2016).

113. Qu, J., Zhou, M., Song, Q., Hong, E. E. & Smith, A. D. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **29**, 2645–2646 (2013).

114. Barros-Silva, D., Marques, C. J., Henrique, R. & Jerónimo, C. Profiling DNA methylation based on next-generation sequencing approaches: New insights and clinical applications. *Genes* (2018) doi:10.3390/genes9090429.

115. Skvortsova, K. *et al.* Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. *Epigenetics Chromatin* **10**, 16 (2017).

116. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* **28**, 1097–1105 (2010).

117. Unnikrishnan, A. *et al.* The role of DNA methylation in epigenetics of aging. *Pharmacol Ther* **195**, 172–185 (2019).

118. Szulwach, K. E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat Neurosci* **14**, 1607–1616 (2011).

119. Chen, H., Dzitoyeva, S. & Manev, H. Effect of aging on 5-hydroxymethylcytosine in the mouse hippocampus. *Restor Neurol Neurosci* **30**, 237–245 (2012).

120. Hadad, N. *et al.* Absence of genomic hypomethylation or regulation of cytosine-modifying enzymes with aging in male and female mice. *Epigenetics Chromatin* **9**, 30 (2016).

121. Stark, R. & Brown, G. DiffBind : differential binding analysis of ChIP-Seq peak data. *Bioconductor* (2011).

122. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243 (2013).

123. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (80-.).* **341**, 1237905 (2013).

124. Wen, L. *et al.* Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol* **15**, R49 (2014).

125. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (80-.).* **336**, 934–937 (2012).

126. Mellen, M., Ayata, P. & Heintz, N. 5-hydroxymethylcytosine accumulation in postmitotic neurons results in functional demethylation of expressed genes. *Proc Natl Acad Sci U S A* **114**, E7812–E7821 (2017).

127. Li, X., Liu, Y., Salz, T., Hansen, K. D. & Feinberg, A. Whole-genome analysis of the

methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res* **26**, 1730–1741 (2016).

128. Wang, L. *et al.* Programming and inheritance of parental DNA methylomes in mammals. *Cell* **157**, 979–991 (2014).

129. Ma, Q., Lu, H., Xu, Z., Zhou, Y. & Ci, W. Mouse olfactory bulb methylome and hydroxymethylome maps reveal noncanonical active turnover of DNA methylation. *Epigenetics* **12**, 708–714 (2017).

130. Kozlenkov, A. *et al.* A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Sci Adv* **4**, eaau6190 (2018).

131. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18**, 67 (2017).

132. Spijker, S. Dissection of rodent brain regions. *Neuromethods* (2011) doi:10.1007/978-1-61779-111-6_2.

133. Li, D., Zhang, B., Xing, X. & Wang, T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* **72**, 29–40 (2015).

134. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp324.

135. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

136. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* (2013) doi:10.1093/biostatistics/kxs033.

137. Jang, H. S. *et al.* Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet* **51**, 611–617 (2019).

138. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

139. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).