Computational and Structural Biotechnology Journal 17 (2019) 507–515

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Recursive Motif Analyses Identify Brain Epigenetic Transcription Regulatory Modules

Sharmi Banerjee [a,b], Xiaoran Wei [b,c], Hehuang Xie [b,c,d,*]

[a] Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061, USA
[b] Biocomplexity Institute of Virginia Tech, Blacksburg, VA 24061, USA
[c] Department of Biomedical Sciences and Pathobiology, Virginia-Maryland College of Veterinary Medicine, Blacksburg, VA 24061, USA
[d] School of Neuroscience, Blacksburg, VA 24061, USA

ARTICLE INFO

Article history:
Received 10 January 2019
Received in revised form 12 March 2019
Accepted 3 April 2019
Available online 9 April 2019

Keywords:
DNA methylation
Transcription factor
Motif
Brain development
Regulatory module
Gene expression

ABSTRACT

DNA methylation is an epigenetic modification modulating the structure of DNA molecule and the interactions with its binding proteins. Accumulating large-scale methylation data motivates the development of analytic tools to facilitate methylome data mining. One critical phenomenon associated with dynamic DNA methylation is the altered DNA binding affinity of transcription factors, which plays key roles in gene expression regulation. In this study, we conceived an algorithm to predict epigenetic regulatory modules through recursive motif analyses on differentially methylated loci. A two-step procedure was implemented to first group differentially methylated loci into clusters according to their correlations in methylation profiles and then to repeatedly identify the transcription factor binding motifs significantly enriched in each cluster. We applied this tool on methylome datasets generated for mouse brains which have a lack of DNA demethylation enzymes TET1 or TET2. Compared with wild type control, the differentially methylated CpG sites identified in TET1 knockout mouse brains differed significantly from those determined for TET2 knockout. Transcription factors with zinc finger DNA binding domains including Egr1, Zic3, and Zeb1 were predicted to be associated with TET1 mediated brain methylome programming, while Lhx family members with Homeobox domains were predicted to be associated with TET2 function. Interestingly, genomic loci from a co-methylated cluster often host motifs for transcription factors sharing the same DNA binding domains. Altogether, our study provided a systematic approach for epigenetic regulatory module identification and will help throw light on the interplay of DNA methylation and transcription factors.

1. Introduction

In mammalian genomes, DNA methylation primarily occurs on the cytosines in the context of CpG dinucleotides and plays an important role in many biological processes including tissue development and cellular function. DNA methylation contributes to the dynamics of chromatin conformation, and thus interferes directly or indirectly with the interactions between DNA molecule and its binding proteins. For instance, the methylation on gene promoters may prevent the bindings of transcription activators but facilitate the recruitment of transcription suppressors [1,2]. In the past decade, high-throughput sequencing technologies and methylation microarray such as the Infinium MethylationEPIC BeadChip provide powerful strategies to generate a large amount of methylome data. Excellent algorithms have been developed for bisulfite sequencing data mapping, array data processing and the identification of differentially methylated CpG sites [3–6]. Despite a growing list of tools, very limited approaches have been developed to predict protein-DNA interaction networks associated with methylation alterations.

Transcription factors (TFs) interpret genetic sequence and epigenetic modification information simultaneously. These proteins recognize specific sequence motifs or structural features of specific genomic regions. To determine the binding sites for transcription factors, ChIP-seq (chromatin immunoprecipitation followed by sequencing) has been widely used. The ChIP-seq technique uses an antibody against a specific TF to pull down the genomic DNA bounded by the target TF for high-throughput sequencing. From ChIP-seq data, *de novo* motif discovery can be achieved with motif discovery tools such as HOMER [7] (Hypergeometric Optimization of Motif EnRichment). HOMER also compiles a database collecting TF motifs from (1) motif database generated by HOMER software largely based on motif analyses using high-quality public ChIP-seq datasets. (2) JASPAR motif database primarily based on published binding site selection experiments in addition to data from *in vitro* microarray-based factor affinity experiments, (3) Organism-centric motif databases including *Drosophila* [8], *Arabidopsis* [9] and *Saccharomyces cerevisiae* motifs [10,11]. Recent

* Corresponding author at: Biocomplexity Institute of Virginia Tech, Blacksburg, VA 24061, USA
E-mail address: davidxie@vt.edu (H. Xie).

https://doi.org/10.1016/j.csbj.2019.04.003

years, several studies have been conducted to explore the transcription factor binding preference on free DNA, nucleosomal DNA and methylated DNA [12,13]. To assess the protein binding affinity to methylated DNA, methylation-sensitive SELEX (systematic evolution of ligands by exponential enrichment) approach has been developed [7]. Transcription factors were incubated with a pool of random methylated and/or unmethylated oligonucleotides to enrich for the ones bounded by the TF of interest. It was found that DNA methylation affects the binding of approximately 60% of the 574 human transcription factors assessed [7]. Currently, such important information has not been fully explored to understand the epigenetic changes observed during normal development and disease progression.

The changes in DNA methylation require the participation of epigenetic machinery including DNA methyltransferases and DNA demethylation enzymes. Ten-eleven translocation (TET) enzymes promote the removal of methyl-group *via* the oxidation of 5-methylcytosines [14]. Three TET family members in mammals were found to have distinct expression patterns and functional activities in different developmental stages and diverse cell types. In mouse brain, *Tet1* is involved in both neural development related biological processes [15] and neuronal activity induced methylation changes [16]. The loss of either *Tet1* or *Tet2* impairs hippocampus neurogenesis and cognition [15,17]. The deletion of *Tet3* leads to neonatal lethality due to its key role in fertilized oocytes to demethylate both maternal and paternal DNA by coupling with DNA replication [18,19]. Despite the growing information for the critical functions of three TET enzymes, it is not yet clear exactly how TET enzymes are recruited to specific genomic loci and the interplays between TET enzymes and transcription factors remain largely elusive.

In this study, we aim to develop a computational pipeline to expedite the identification of epigenetic transcription regulatory modules (ETRMs) from differentially methylated genomic loci. Here, an ETRM refers to a set of TFs with binding sites adjacent to a 'key' TF, whose motif is the most significantly enriched in differentially methylated regions. To demonstrate the power of our analytical procedure, we made use of methylome data sets generated for the frontal cortices from TET1 knockout (TET1KO) and TET2 knockout (TET2KO) mice and predicted that several neural-development-related or neuronal-activity-induced ETRMs were associated with brain methylome programming on different genomic loci mediated by TET1 or TET2. The software package developed in the study is available in Github repository ETRM-identification (https://github.com/BSharmi/ETRM-identification).

2. Results

2.1. An Outline of the Computational Pipeline for ETRM Identification

Most methylation studies led to the identification of CpG sites or genomic regions showing differential methylation patterns among various conditions. The association of methylation changes to a specific physiological condition or disease statuses could be established swiftly, however, the understanding of mechanistic causal nexus underlying dynamic methylation often requires substantial extra effort. The determination of transcription factors as methylation readers and effectors is crucial to explain how methylation changes occur and translate into the alterations of downstream gene expression. To ease this process, we conceived a two-step approach (Fig. 1, Supplementary Fig. S1) to take differentially methylated regions (DMRs) as inputs: 1) to partition DMRs into co-methylated clusters according to their methylation correlations using the Weighted Correlation Network Analysis (WGCNA) algorithm [20]; 2) to determine TF motifs enriched in each cluster using a recursive motif-finding algorithm.

The first step in our approach is to identify co-methylated clusters following a procedure described in our previous study [21], which is based on the assumption that: genomic loci sharing similar methylation profiles during development or among cell types might be co-regulated by a common set of TFs. Since the subsequent motif enrichment analysis will be performed using a hyper-geometric test to compare input sequences with background controls, the motif enrichment *p*-values depend heavily on the composition of input sequences. This first step greatly reduces genome complexity *via* segmentation according to the similarity in methylation, and thus increases the chances to identify motifs with significant *p*-values. In the second step, we implemented a novel recursive motif search algorithm to identify transcriptional regulatory modules. For each co-methylated cluster, we performed motif enrichment analysis using HOMER to identify the transcription factor with a motif enriched with the most significant *p*-value, and denoted as "key TF". From the co-methylated cluster, genomic regions contain the motif of the key TF will be extracted as a sub-cluster for another round of motif search. For this sub-cluster, we will be able to determine the transcription factors with motifs significantly enriched in genomic regions surrounding the motif of the key TF. This step establishes the links between key TF and its associated TFs for a sub-cluster, which comprised differentially methylated regions sharing a similar methylation profile. Such a process will be recursively executed until no motif could be identified with an enrichment *p*-value less than 1e-10 as recommended by the HOMER software. Therefore, each co-methylated cluster may be further divided into several sub-clusters. For each sub-cluster, an epigenetic regulatory module with a key TF and associated TFs will be predicted. Further experimental data from ChIP-seq studies could lend a hand with supporting evidence for the ETRMs predicted.

2.2. A Comprehensive Motif Database Compiled for Epigenetic Regulatory Module Identification

In this study, we aimed at identifying epigenetic regulatory modules, particularly for the transcription factors with motifs enriched within differentially methylated regions. To achieve this goal, we augmented the HOMER's database of known motifs with two additional motif datasets (Fig. 2). The first is a combined set of un-methylated (or canonical) and methylation-related motifs compiled in the MeDReaders database according to published literature [22]. This dataset contains motifs for 731 transcription factors (601 for human and 130 for mouse) that may bind to methylated DNA. In addition, the MeDReaders database also provides methylation-related motifs predicted using *in silico* approaches for 292 transcription factors (287 for human and 5 for mouse). In addition to this motif library, we extended the search to include information regarding how DNA methylation may affect TF binding. A second dataset was prepared using the position frequency matrices (PFMs) for over five hundred transcription factors obtained with methylation-sensitive SELEX approach [12]. The motifs in this dataset can be broadly classified as (a) 'MethylMinus': the consensus sequence obtained from the motif with one or more CGs, the methylation of which negatively affects TF binding; (b) 'MethylPlus': the consensus sequence obtained from the motif contains one or more CGs, the methylation of which enhances TF binding. Thus, we integrated the aforementioned two motif datasets with HOMER known motif database and classified the motifs into five types – (1) methylation-related motifs, containing motifs without CpG dinucleotide ('No CpG'), motifs with CpG but methylation having little effect on TF binding ('Little effect') [12] and motifs predicted in methylated sequences by MeDReaders using *in silico* approaches (2) canonical motifs predicted from un-methylated sequences by MeDReaders; (3) 'MethylPlus' and (4) 'MethylMinus' identified with methylation-sensitive SELEX approach; and (5) motifs from HOMER database.

After PFM deduplication, we collected a total of 364 motifs from the HOMER database, 864 canonical motifs from MeDReaders, 22 non-canonical motifs from MeDReaders, 191 'MethylPlus' motifs and 143 'MethylMinus' motifs. The transcription factors associated with these motifs were summarized in Fig. 2A. The TFs documented in our integrated motif database can be classified according to their DNA-binding domains [23] and shown in Fig. 2B using a Circos plot [24]. We observed that 30.1% of motifs documented are for TFs with

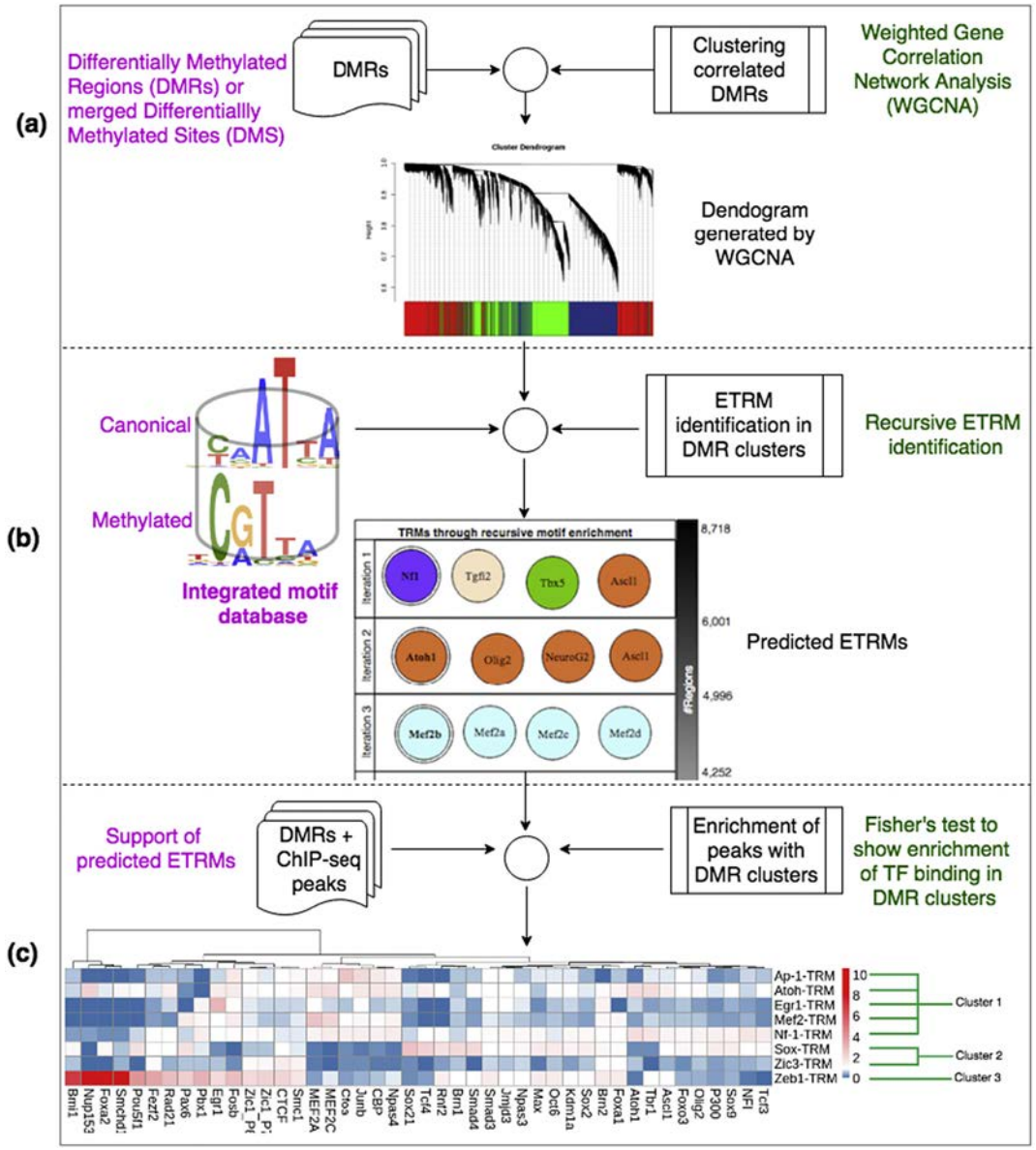

Fig. 1. Computational pipeline designed to identify transcriptional regulatory modules (ETRMs) within differentially methylated regions using recursive motif searching algorithm. The pipeline is comprised of three segments: (a) WGCNA is applied on the methylation profiles of differentially methylated regions (DMR) to obtain co-methylated clusters. (b) A recursive algorithm (Supplementary Fig. S1) is used to identify ETRMs for each co-methylated clusters. Color codes were used to indicate TF families. (c) TF ChIP-seq datasets are exploited to support the predicted ETRMs. The left column shows the input data and the method used and the right column illustrates examples for output files.

Homeobox domain whereas only 0.2% of motifs were for TFs with the MAD domain. Transcription factors with the homeodomain (HM) and C2H2 Zinc Fingers (ZF) domain frequently recognize the motifs in the 'MethylPlus', 'Canocial' and 'Methylated' categories. On the other hand, transcription factors with high mobility group box (HMG-box) and E26 transformation-specific (ETS) domain were present in all except for the 'MethylPlus' category. Although the information on methylation preference is still incomplete for many TFs, this result suggests that the differences in DNA binding domains could be a key factor controlling whether a TF would interact with their methylated binding sites.

2.3. Genome Distribution of Hypermethylated CpG Sites Identified in TET1KO and TET2KO Frontal Cortices

With this unique motif database, we applied the analytical pipeline implemented (Fig. 1) on two methylome datasets generated for the frontal cortices of TET1KO and TET2KO mice. The TET1KO methylome was generated with reduced representation bisulfite sequencing to enrich for genomic regions rich in CpG dinucleotides by digestion with *Mse*I and *Mlu*CI enzymes with recognition sites for TTAA and AATT, respectively. The TET2KO methylome was generated with whole genome bisulfite sequencing, and thus with broader coverage in genome-wide. Interestingly, we identified 42,558 differentially methylated sites (DMSs) for TET1KO but only 12,900 DMSs for TET2KO. In addition, only 643 common DMSs were identified between TET1KO and TET2KO tissues (Supplementary Fig. S2A). This result suggests TET1 and TET2 enzymes are indispensable for two distinct sets of CpG sites during brain development. Compared to the genome distribution of TET2KO DMSs, TET1KO DMSs tend to localize inside or adjacent to genes (Supplementary Fig. S2B&C). For instance, promoters host 4.3% of TET1KO DMSs and 1.9% of TET2KO DMS respectively. Thus, promoter methylation seems more susceptible to the loss of TET1 than the loss of TET2. Since both TET1 and TET2 are DNA demethylation enzymes, the direct consequence of TET1 or TET2 loss is the increased methylation on their corresponding targets. It is not a surprise that we found the majority of DMSs identified (75.1% for TET1KO and 67.3% for TET2KO) are

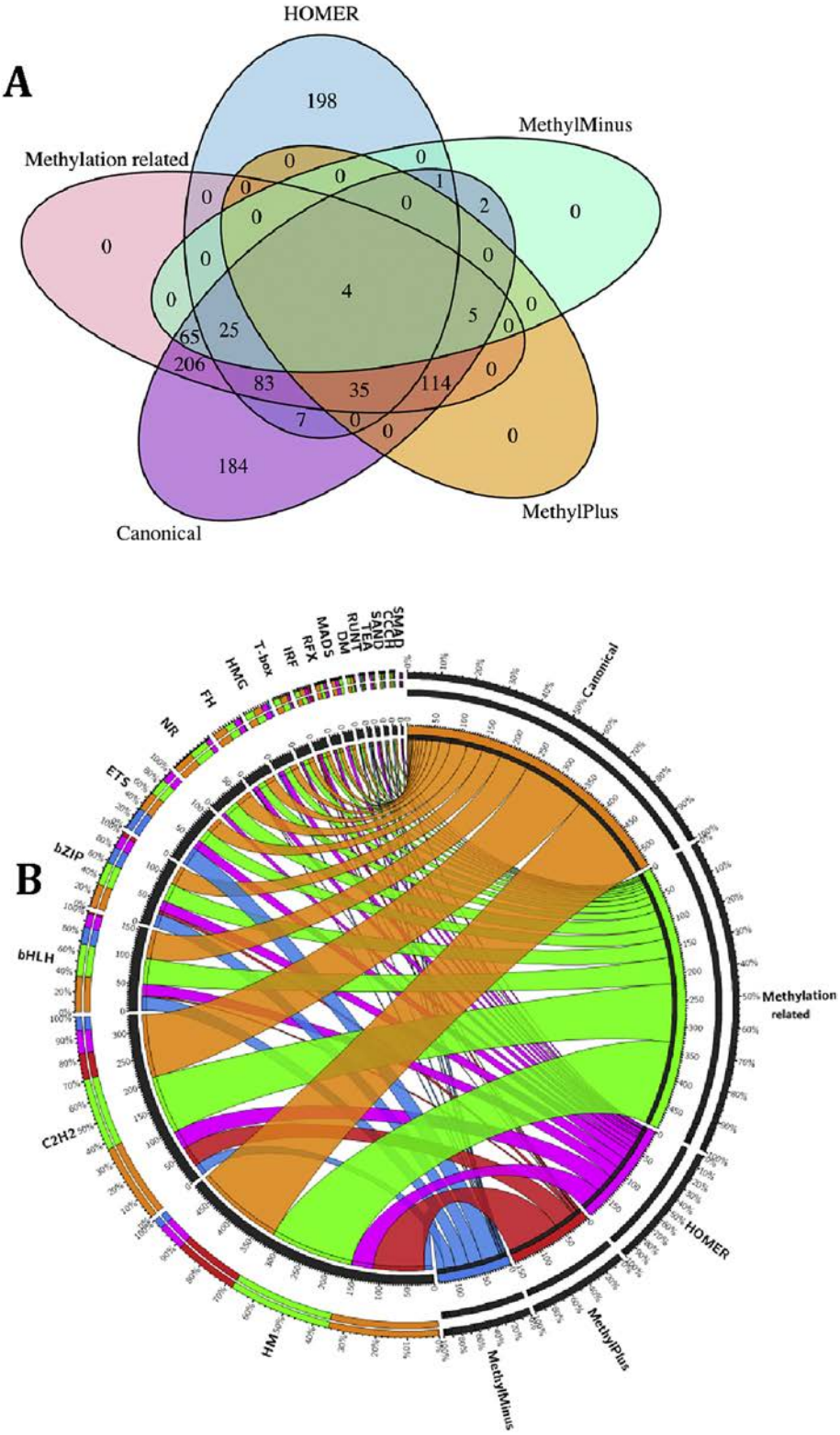

Fig. 2. The composition of methylation related motif database. A) Venn diagram showing the number of shared motifs among the five categories. B) Distribution of five motif categories: Canonical, Methylation-related, HOMER, 'MethylPlus' and 'MethylMinus'. TF binding domains were arranged in clockwise decreasing order.

hypermethylated in TET1KO or TET2KO mice. On the other hand, the methylation losses observed in TET1KO or TET2KO mice are likely to be indirect consequences. In order to understand the gain of methylation observed in TET1KO and TET2KO mice, further analyses were limited to the hypermethylated regions surrounding DMSs identified in TET1KO and TET2KO mice.

2.4. Distinct ETRMs Identified for Differentially Methylated Clusters

To perform co-methylation co-regulation analysis, we collected 66 methylomes for mouse forebrain, midbrain and hindbrain regions during development and for sorted mouse brain cells including astrocytes, oligodendrocytes, Excitatory neurons, PV neurons and VIP neurons

(Supplementary Table S1A). We next applied WGCNA to cluster differentially methylation regions into groups, which show highly correlated methylation profiles across the methylomes. For TET1KO or TET2KO, the WGCNA algorithm identified three co-methylated clusters with each cluster showing a distinct methylation profile (Figs. 3 & 4, Supplementary Fig. S3). Despite TET1KO and TET2KO mice share a very small number of DMSs as noted earlier, we observed similar methylation profiles for some clusters in two kinds of mice of distinct genotypes. For instance, the first clusters in TET1KO and in TET2KO both show decreased methylation during brain development and are with the lowest methylation level in excitatory neurons among all brain cell types. For TET1KO, the second cluster shows low methylation during embryonic development phases and increased methylation in postnatal frontal cortex while the third cluster shows hypo-methylation in most samples except in TET1KO. For TET2KO, the second cluster shows increased methylation in various kinds of neurons *vs* glial cells while the third cluster shows hyper-methylation in astrocyte, E17.5 neuron, and oligodendrocyte.

For each cluster, we applied HOMER software to identify transcription factor motifs within the sequences of differentially methylated loci. Top significant motifs were shown in Figs. 3 & 4 for each ETRM with more details in Supplementary Tables S2. Despite the fact that the majority of DMSs are different in TET1KO and TET2KO mice, several transcription factors were identified to be associated with both TET1KO and TET2KO including MEF2 family with MADS domain (MCM1, Agamous, Deficiens, and Serum response factor), Sox family with HMG domain (high mobility group) and Olig2, Ascl1 from bHLH domain (basic helix-loop-helix). Interestingly, for TET1KO, transcription factors from Zinc Finger family were enriched in all three clusters but the cluster 1 showed motif enrichment for EGR1 whereas cluster 2 and 3 showed the motif enrichment for ZIC3 and Zeb1, respectively (Supplementary Tables S2). EGR1 is involved in the consolidation of new memories and critical for early postnatal brain development [25]. ZIC3 regulates the expansion of neuronal progenitors [26] while Zeb1 keeps neurons in an immature state by preventing neuronal polarization [27]. This result indicates that TET1KO associated TFs with the same DNA binding domain may play very different functional roles during brain development. It helps in explaining why these ZF TFs with motifs enriched in different genomic regions showing distinct methylation profiles.

We next applied the recursive motif identification approach on each cluster to investigate motif distribution. During the recursive motif searching process, the exclusion of some genomic regions in a cluster may lead to substantial changes in the enrichment *p*-values for TF motifs. For example, in cluster 1 of TET1KO, Egr2, Sp5, and Klf14 motifs are with enrichment *p*-values as 1e-5, 1e-3 and 1e-6 respectively (Supplementary Tables S2A). However, all three motifs were significantly enriched in Egr1 ETRM, a subset in the cluster1, which was obtained through a recursive motif identification approach (Fig. 3D). Thus, the recursive ETRM identification approach predicted clusters of motifs enriched nearby a key TF, which otherwise go unnoticed if all input sequences are analyzed together. The recursive motif identification approach also led to several interesting observations. Some TFs with different DNA binding domains are likely to form an ETRM together, such as NF1 (CTF domain), Tgfi2 (Homeobox domain) and Ascl1 (bHLH domain) (Figs. 3D & 4D). On the other hand, some ETRMs are composed of TFs with DNA binding domain of the same class, including TFs with the ZF domain (EGR family), MADS domain (MEF2) and HMG domain (SOX family). These results shed some light on how TFs may be organized into regulatory complex, for instance, heterodimers with TFs from the same family. Pioneer TFs such as Ascl1 [28] were observed to have motifs enriched in multiple ETRMs (Fig. 3D). This may reflect the fact that pioneer TFs act at early stages and participate in epigenetic regulation of

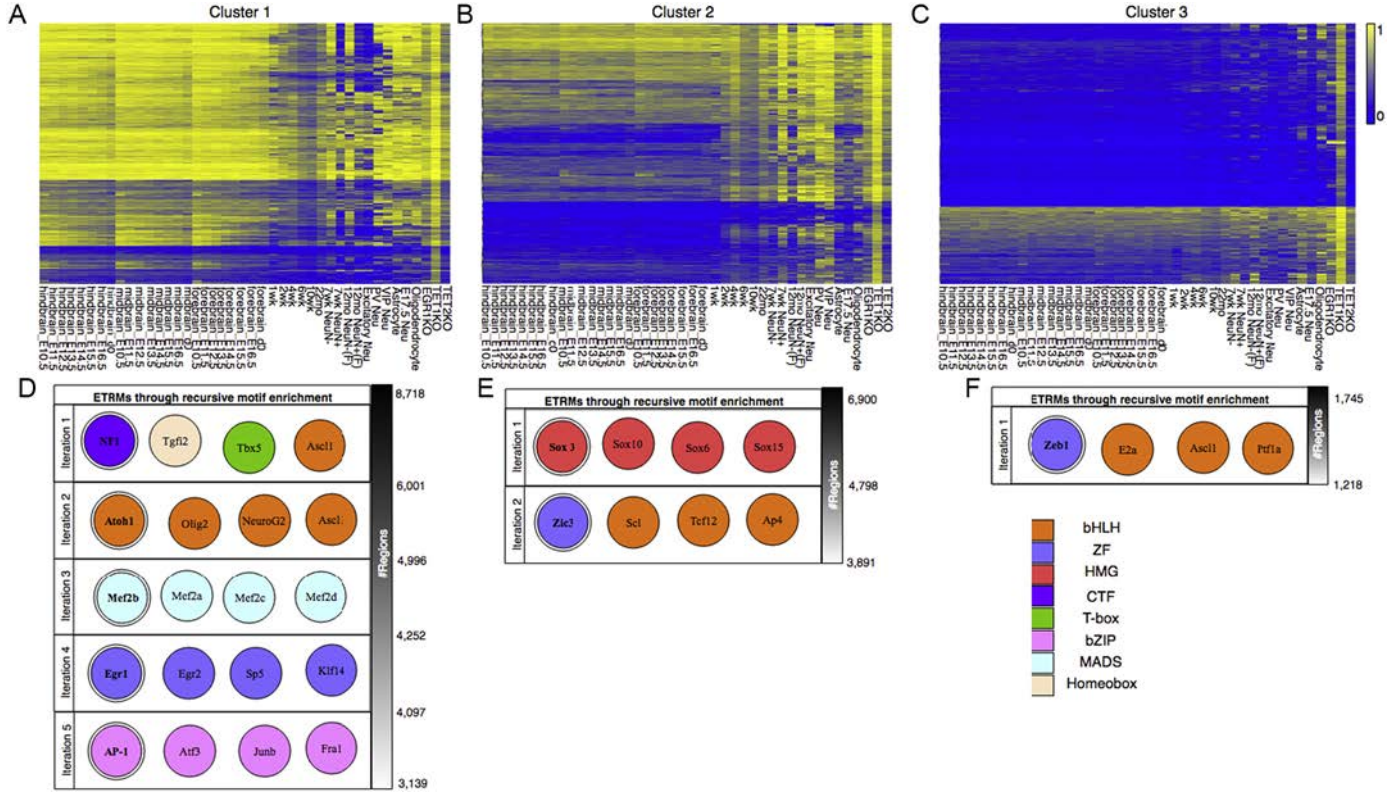

Fig. 3. Methylation profiles of three DMR clusters identified (A-C) and the corresponding ETRMs predicted (D-F) for TET1KO brain methylome. The range of methylation levels was set as 0 (blue) to 1 (yellow). For each iteration cycle, the numbers of DMR included in the analysis were shown on right bar. The key TF identified in each iteration cycle was marked with double ring. The TFs were colored according to TF family annotated with DNA binding domain.

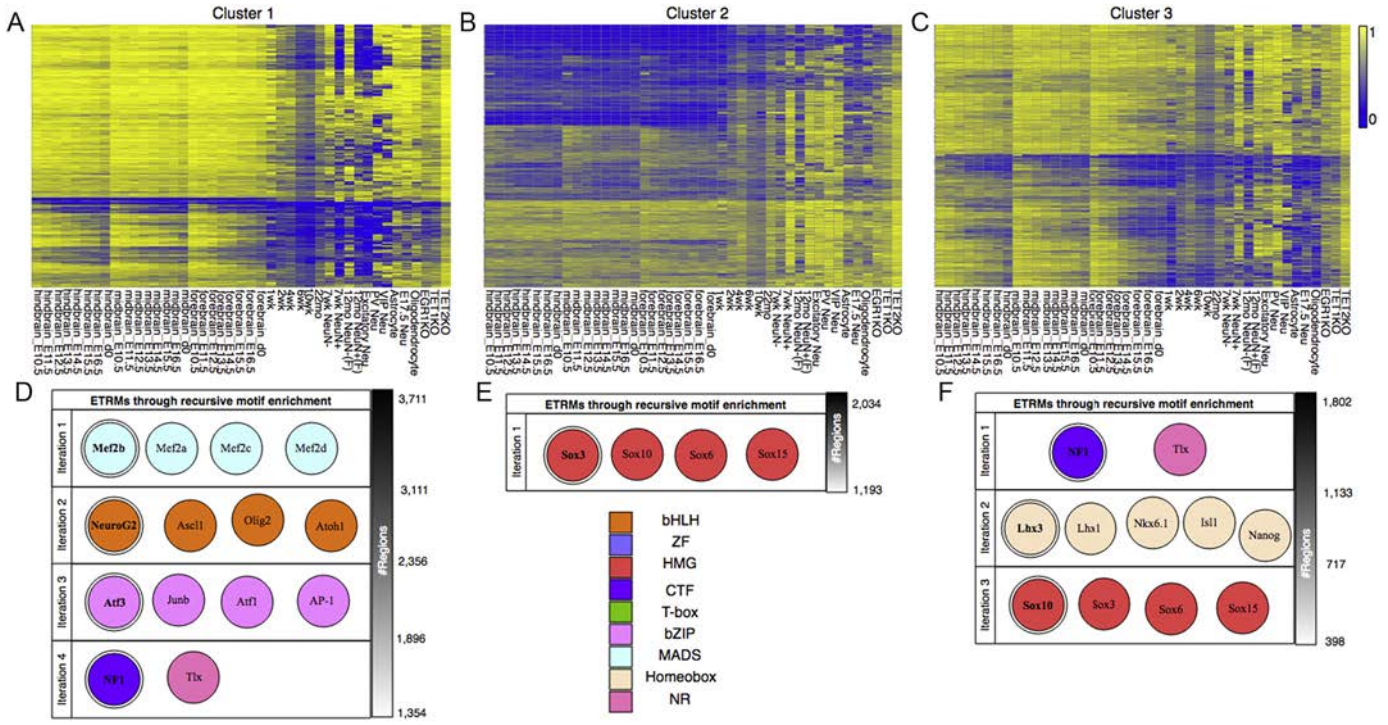

Fig. 4. Methylation profiles of three DMR clusters identified (A-C) and the corresponding ETRMs predicted (D-F) for TET2KO brain methylome. The range of methylation levels was set as 0 (blue) to 1 (yellow). For each iteration cycle, the numbers of DMR included in the analysis were shown on right bar. The key TF identified in each iteration cycle was marked with double ring. The TFs were colored according to TF family annotated with DNA binding domain.

genomic regions in multiple sub-clusters. Interestingly, Sox3 and Sox10 were predicted to be key TFs in TET2KO clusters. The transient expression of Sox3 has been reported in neural progenitors [29] but Sox10 is known as a critical regulator involved in multiple stages during neural crest development [30]. The genomic regions related to Sox3 ETRMs are hypo-methylated whereas those for Sox10 ETRM are hyper-methylated during development (Supplementary Fig. S4).

Transcription factors binding could be under the influence of DNA methylation, histone modifications, and chromatin structures [12,31]. For ETRMs identified for both TET1KO and TET2KO brains, several methylation-related motifs were enriched in the differentially methylated regions (Supplementary Table S3). As classified in our motif database, motifs from following categories were identified for their corresponding TFs: (1) 'MethylMinus' motifs for Sox12, Atf3, Batf and Tcf12; (2) 'MethylPlus' motifs for Lhx1, Lhx4, Lhx6, Lhx9 and Nanog; (3) 'Methylation-related' motifs, which include 'motifs with little effect by DNA methylation' for Tlx2 and Myog and 'motifs with no CpG' for Atoh1, MEF2 and Sox family members (Supplementary Fig. S5). Having 'motifs with little effect by DNA methylation' or 'motifs with no CpG' would enable a TF to neglect methylation statuses of its binding sites. Such kind of TFs may work as forerunners together with TFs bind to 'MethylPlus' motifs. Logically, the bindings of TFs with 'MethylMinus' motifs may depend on the interactions between forerunners and DNA demethylation enzymes to demethylate their binding sites. Thus, the identification of ETRMs together with the information of methylation related motifs would supply information for how TFs may assemble in differentially methylated loci.

2.5. Brain ChIP-seq Data Support the ETRMs Predicted

We next made use of brain ChIP-seq data to determine the binding frequencies of transcription factors on the genomic sequences for each differentially methylated sub-cluster with ETRM identified (Fig. 5). Not surprisingly, most ETRMs we predicted are enriched for the binding sites of corresponding transcription factors. For instance, the top TFs with binding sites enriched in AP-1-ETRM are FOS and JUN proteins, the two sub-units of a typical AP-1 protein complex. Similar results were observed for Atoh-ETRM, Egr1-ETRM, Mef2-ETRM, NF1-ETRM, and Sox-ETRM. Although we did not find the ChIP-seq study for Zic3 in mouse brain, genome-wide analysis of Zic3 binding sites in zebrafish embryos revealed a distribution biased towards distal intergenic regions that may act as functional enhancers [32]. We observed that Zic3-ETRM is enriched for the binding sites of chromatin domain boundary proteins CTCF and its partner Smc1 (Structural maintenance of chromosomes protein 1) which is a key component of cohesin ring complex [34]. Four transcription factors are with binding sites depleted from Zic3-ETRM but highly enriched in Zeb1-ETRM are Nup153, Foxa2, Bmi1, and Smchd1 (Fig. 5A). Although the interactions among Zeb1 with these transcription factors are not well studied, Zeb1, Foxa2, Bmi1, and Smchd1 were reported to be critical for epithelial mesenchymal transition [27,35–37], an early neural developmental process in which the neuroepithelium of the dorsal neural tube gives rise to neural crest cells.

For both TET1KO and TET2KO methylomes, we observed neuronal-activity-related transcription factors Npas4 and co-factor CBP are with binding sites clustered together in AP-1-ETRM and Mef2-ETRM (Fig. 5A). On the other hand, brain-development-related transcription factors Pax6, Tcf4, and Brn1 are enriched in Lhx-ETRM and Sox-ETRM (Fig. 5B). Mef2-ETRM, NF1-ETRM, and Sox-ETRM were identified for both TET1KO and TET2KO DMRs. Despite the similarity in the enrichment of corresponding transcription factors, a number of TFs show distinct binding preferences for these three ETRMs identified in the mouse brains lacking the two different TET enzymes. In particularly, Nup153 and Smchd1 binding sites are highly enriched in regions regulated by Mef2-ETRM in Tet2KO but depleted from those in the Tet1KO brain. Currently, little information is available regarding the interactions between TET1 and TET2 enzymes with transcription factors such as MEF2 family members, our result suggests TET1 and TET2 are likely involved in two independent epigenetic regulatory pathways, *i.e. via* different combinations of TFs with MEF2.

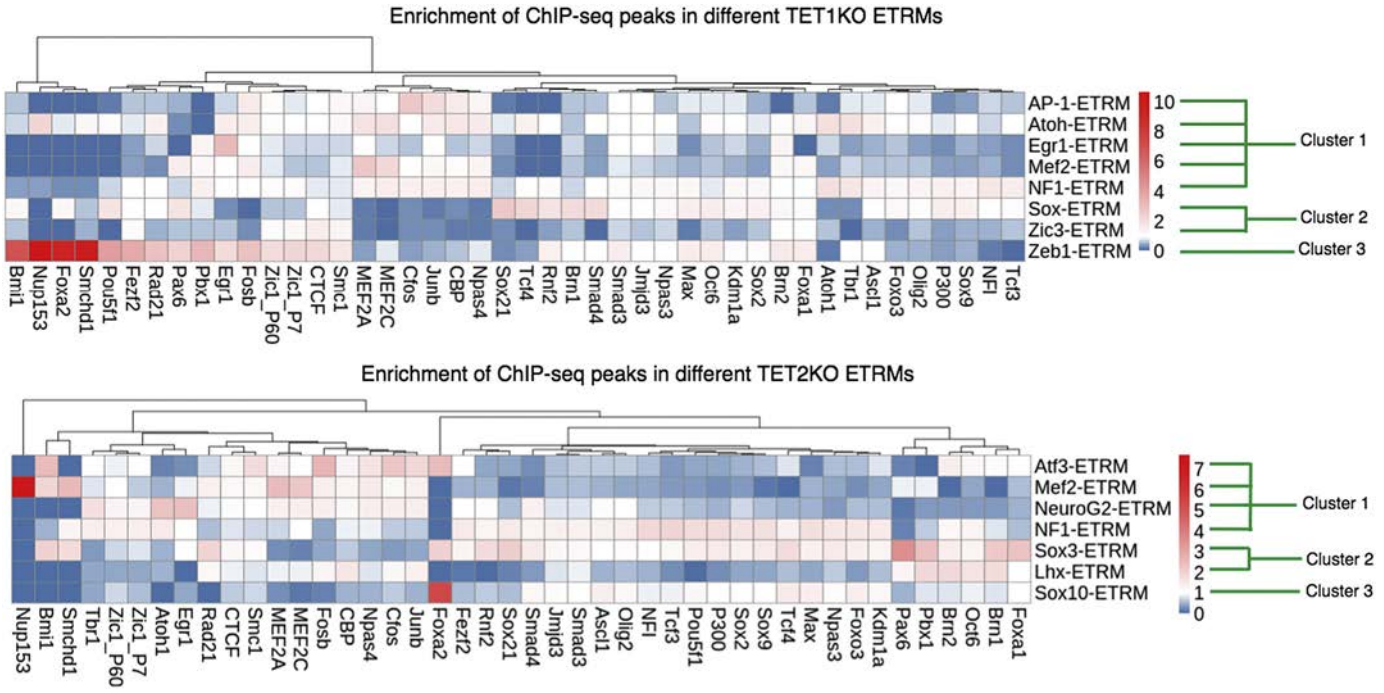

Fig. 5. The enrichment of TF binding sites in TET1KO and TET2KO ETRMs. Each color value in the figures represents the estimate of the odds ratio based on conditional Maximum Likelihood Estimate derived from Fisher's exact test. For TET2 KO, the two NF1 ETRMs identified from cluster 1 and cluster 3 shown in Fig. 4 were combined and labeled as one NF1-ETRM in the bottom panel.

3. Discussion

In this study, we proposed a computational pipeline to maximize the mechanistic understanding of methylation changes and demonstrated its applications with brain methylomes derived from TET1KO and TET2KO mice. To our knowledge, this is the first toolkit taking advantage of several lines of information embedded in existing datasets: 1) methylome datasets collected for DMR clustering; 2) motif database compiled for methylation associated TF searching; and 3) ChIP-seq datasets providing additional experimental evidence.

From technical aspects, our approach consisted of two complementary steps: clustering based on methylation correlation and recursive motif identification based on sequence analysis. These two steps serve an important purpose to group genomic loci under the same epigenetic regulation mechanism together and thus improve the likelihood to identify TFs with motifs significantly enriched in each sub-group. Our pipeline accepts DMRs identified with different thresholds defined by various kinds of software designed for methylation data analysis. Worthy of mention, ETRMs identified with our pipeline are not based on a single DMR but rather thousands of DMRs sharing a similar methylation profile. Therefore, it is not difficult to imagine that a slight change in the list of DMRs is unlikely to result in striking differences in ETRMs predicted. More accurate clustering results in grouping DMRs sharing similar methylation patterns may be achieved with the increasing numbers of methylomes generated for diverse conditions, development stages and distinct cell types. The recursive TF motif identification enables us to further explore the differences within clusters at the sequence level and to reveal diverse mechanisms driving the epigenetic dynamics. The combination of these two steps also allows us to explore subtle differences in epigenome regulation within a specific cell type at a given developmental stage. Our pipeline adopted the WGCNA algorithm to group genomic regions with highly correlated methylation profiles. Weighted adjacency feature in WGCNA puts emphasis on regions with a high correlation at the expense of regions with low correlations where the weight can be selected using scale-free topology [38]. WGCNA also allows scalability by splitting large matrices into smaller blocks to fit within the available RAM resulting in faster computation. For differentially methylated regions, we recommended a window size comparable to the width of a typical ChIP-seq peak. In a recent study conducted on defining features for ChIP-seq peak calling algorithms, the authors suggested the ChIP-seq peak as a 200 bp window surrounding the peak center [39]. They observed that the performance of some peak calling tools dropped with a setting for shorter window width (75 bp per window). Finally, HOMER motif analysis is limited to a set of TFs with known motif position weight matrix (PWM). For instance, based on ChIP-seq data, we observed the enrichment of Nup153, Bmi1, and Smchd1 in cluster 3 of TET1KO. However, in HOMER or JASPAR [40] databases, no motif is documented for these three TFs. In addition, some transcription factors prefer to recognize DNA structure (shape-based) instead of a stretch of DNA sequences, and thus not all TFs are with motifs documented in the motif databases.

From biological aspects, we identified a number of brain ETRMs composed of transcription factors from the same family. This result is consistent with known facts that some TFs, such as MEF2 [41] and AP-1 [42] (a heterodimer composed of Fos and Jun proteins), tend to form homo-dimers with itself or heterodimers with members of the same family. Worthy of mentioning, TFs from the same family may not work on the same locus simultaneously but substitute each other in a sequential order during the developmental process. For instance, Sox2 and Sox3 keep neuronal differentiation genes silent in neural progenitor cells and will be replaced by Sox11 when terminal differentiation initiates [43]. Apparently, the lack of comprehensive datasets of high-quality could dampen the power of the pipeline to provide an accurate prediction. For instance, brain ChIP-seq datasets used in this study were generated from diverse neural stem cell lines and brain tissues of different developmental stages. Currently, very limited co-binding information is available for these transcription factors due to the lack of ChIP-seq datasets generated on desired experimental conditions. Additionally, ETRMs were identified from different sets of genomic loci but may interact with each other in 3D through chromatin folding. Future efforts are required to integrate chromatin configuration information into ETRM modeling. Despite these limitations, the predicted ETRMs could still provide great starting points for further dedicated research and we anticipate the analytic procedure described in this study will

assist in the ultimate interpretation of the causes and consequences of methylation alterations. Finally, the combination of HOMER with the recursive motif search algorithm for regulatory module identification may be applied on genomic regions of other interests beyond differentially methylated loci demonstrated in this study.

4. Materials & Methods

4.1. "Omics" Datasets and Data Processing

Methylome and ChIP-seq datasets used in the study are summarized in Supplementary Table S1A & B, respectively. Methylome datasets were processed as described in previous studies with slight modifications [44,45]. Sequence bases of low quality and illumina adaptors were trimmed off using Trim_Galore. Trimmed sequences were aligned to mouse reference genome mm10 using Bismark [3]. Fisher Exact test was used to evaluate the significance of differential methylation [45]. Briefly, a contingency table was constructed for each CpG with the rows indicated two conditions and the columns indicated the number of methylated cytosines and unmethylated cytosines. In the test, CpG sites were required to have at least 10Xs read coverage. A sequential permutation method was employed to control FDR [46]. A new contingency table was reconstructed by randomly assigning reads to cells with the same methylation probability for each sample in each permutation. A total of 1000 permutations were performed for each CpG site. Differentially methylated sites were determined with an adjusted *p*-value equal to or less than 0.05. Neighboring DMSs located within 200 bp window were merged into DMRs. Finally, all DMRs including orphan DMSs (in absence of neighboring DMS within 200 bp) were extended by 100 bp to both ends and used as the inputs for recursive ETRM identification. ChIP-seq data processing followed the procedure described in our previous study [31].

4.2. Clustering DMRs for motif Enrichment

WGCNAR package [20] was used to group each DMR set into different clusters. We collected 66 methylomes for mouse forebrain, midbrain and hindbrain regions during development and for sorted mouse brain cells including astrocytes, oligodendrocytes, excitatory neurons, PV neurons and VIP neurons (Supplementary Table S1A). For each DMR, a matrix was generated to host the methylation values from the 66 methylomes. There are two main steps in WGCNA clustering. In the first step, DMRs were pre-clustered into different blocks using projective k-means. Next, for each block, network analysis was performed by identifying clusters of highly correlated DMRs and estimating cluster Eigen node which is the first principal component of a module and can be considered as a representative of the methylation profile in a module. Lastly, clusters with highly correlated Eigen nodes were merged. According to WGCNA clustering, a label '0' (color 'Grey') was assigned to DMRs that were not part of any co-methylated module and were excluded from further analysis.

4.3. Preparing Motif Libraries

HOMER known motif library is primarily based on the analysis of high-quality ChIP-Seq data sets [7]. The two additional methylation-related motif libraries contain position weight matrices (PWM) for individual motifs were added to the HOMER database under known TF motif directory. For annotation, all methylation-related motifs have '_methylated' appended to their names.

4.4. Recursive Motif Search to Identify ETRMs

Known motif enrichment analysis was performed using the script findMotifs.pl in HOMER with parameter "–mset vertebrates". The motif with the most significant *p*-value predicted by HOMER was selected as a key TF. In case of ties involving two motifs sharing a same enrichment *p*-value, the motif with a higher frequency in target sequences was selected. Next, the regions containing the key motif were identified with HOMER and the center of each region was shifted to the predicted binding site of the key TF for another round of motif search. This step results in the identification of TF motifs adjacent to the key motif. Finally, the regions containing the motif for the key TF were removed from the input dataset and the rest of input sequences were used to identify the next most significant candidate motif. Such motif searching process was performed recursively until no significant motif can be identified.

Author Contributions

H.X. conceived and designed the study; S.B. implemented clustering and recursive motif finding procedure; S.B. and X.W. conducted data analysis and integration; S.B. and H.X. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Acknowledgments

This work was supported by NIH grant NS094574 and the Biocomplexity Institute faculty development fund for H.X, and VT's Open Access Subvention Fund.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.04.003.

References

[1] Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell 1991;64:1123–34.
[2] Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet 2016;17:551–65.
[3] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics 2011;27:1571–2.
[4] Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. Genome Biol 2012;13:R61.
[5] Assenov Y, et al. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods 2014;11:1138–40.
[6] Morris TJ, et al. ChAMP: 450k Chip analysis methylation pipeline. Bioinformatics 2014;30:428–30.
[7] Heinz S, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 2010;38:576–89.
[8] Kulakovskiy IV, Favorov AV, Makeev VJ. Motif discovery and motif finding from genome-mapped DNase footprint data. Bioinformatics 2009;25:2318–25.
[9] Steffens NO, Galuschka C, Schindler M, Bulow L, Hehl R. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. Nucleic Acids Res 2004;32:D368–72.
[10] Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431:99–104.
[11] MacIsaac KD, et al. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinform 2006;7.
[12] Yin Y, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science 2017;356.
[13] Zhu F, et al. The interaction landscape between transcription factors and the nucleosome. Nature 2018;562:76–81.
[14] Tahiliani M, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 2009;324:930–5.
[15] Zhang RR, et al. Tet1 regulates adult hippocampal neurogenesis and cognition. Cell Stem Cell 2013;13:237–45.
[16] Guo JU, Su Y, Zhong C, Ming GL, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell 2011;145:423–34.
[17] Gontier G, et al. Tet2 rescues age-related regenerative decline and enhances cognitive function in the adult mouse brain. Cell Rep 2018;22:1974–81.
[18] Guo JU, et al. Neuronal activity modifies the DNA methylation landscape in the adult brain. Nat Neurosci 2011;14:1345–51.

[19] Shen L, et al. Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. Cell Stem Cell 2014;15:459–71.
[20] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform 2008;9(559).
[21] Luo YT, et al. Integrative single-cell omics analyses reveal epigenetic heterogeneity in mouse embryonic stem cells. PLoS Comput Biol 2018;14.
[22] Wang G, et al. MeDReaders: a database for transcription factors that bind to methylated DNA. Nucleic Acids Res 2018;46:D146–51.
[23] Wingender E, Schoeps T, Haubrock M, Krull M, Donitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res 2018;46:D343–7.
[24] Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–45.
[25] Duclot F, Kabbaj M. The role of early growth response 1 (EGR1) in brain plasticity and neuropsychiatric disorders. Front Behav Neurosci 2017;11.
[26] Inoue T, Ota M, Ogawa M, Mikoshiba K, Aruga J. Zic1 and Zic3 regulate medial forebrain development through expansion of neuronal progenitors. J Neurosci 2007;27:5461–73.
[27] Singh S, et al. Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. Elife 2016;5.
[28] Wapinski OL, et al. Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell 2013;155:621–35.
[29] Wang TW, et al. Sox3 expression identifies neural progenitors in persistent neonatal and adult mouse forebrain germinative zones. J Comp Neurol 2006;497:88–100.
[30] Kelsh RN. Sorting out Sox10 functions in neural crest development. Bioessays 2006;28:788–98.
[31] Banerjee S, et al. Identifying transcriptional regulatory modules among different chromatin states in mouse neural stem cells. Front Genet 2019. https://doi.org/10.3389/fgene.2018.00731. eCollection 2018 https://www.ncbi.nlm.nih.gov/pubmed/30697231.
[32] Winata CL, et al. Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish. PLoS Genet 2013;9:e1003852.
[34] Holwerda SJ, de Laat W. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos Trans R Soc Lond B Biol Sci 2013;368:20120369.
[35] Zhang X, Wei C, Li J, Liu J, Qu J. MicroRNA-194 represses glioma cell epithelialtomesenchymal transition by targeting Bmi1. Oncol Rep 2017;37:1593–600.
[36] Costello I, et al. Lhx1 functions together with Otx2, Foxa2, and Ldb1 to govern anterior mesendoderm, node, and midline development. Genes Dev 2015;29:2108–22.
[37] Shaw ND, et al. SMCHD1 mutations associated with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome. Nat Genet 2017;49:238–48.
[38] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mo B 2005;4.
[39] Thomas R, Thomas S, Holloway AK, Pollard KS. Features that define the best ChIP-seq peak calling algorithms. Brief Bioinform 2017;18:441–50.
[40] Mathelier A, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 2014;42:D142–7.
[41] Wu Y, et al. Structure of the MADS-box/MEF2 domain of MEF2A bound to DNA and its implication for myocardin recruitment. J Mol Biol 2010;397:520–33.
[42] Riesgo-Escovar JR, Hafen E. Common and distinct roles of DFos and DJun during Drosophila development. Science 1997;278:669–72.
[43] Sarkar A, Hochedlinger K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. Cell Stem Cell 2013;12:15–30.
[44] Zhao L, et al. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. Genome Res 2014;24:1296–307.
[45] Lister R, et al. Global epigenomic reconfiguration during mammalian brain development. Science 2013;341:1237905.
[46] Bancroft T, Du C, Nettleton D. Estimation of false discovery rate using sequential permutation p-values. Biometrics 2013;69:1–7.