Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Tao Wang Date

TET-mediated Hydroxymethylation in Reprogramming to Induced Pluripotency

By

Tao Wang (王韬)

Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

_____

Stephen T. Warren
Co-Advisor

_____

Peng Jin
Co-Advisor

_____

Victor Corces
Committee Member

_____

Hao Wu
Committee Member

_____

Anthony Chan
Committee Member

_____

Stephen Dalton
Committee Member
Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

TET-mediated Hydroxymethylation in Reprogramming to Induced Pluripotency

By

Tao Wang (王韬)

B.S., University of Science and Technology of China, Hefei, China 2005

Advisor: Stephen T. Warren, Ph.D.

Peng Jin, Ph.D.

An abstract of
A dissertation submitted to the Faculty of
the James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

2013

Abstract

TET-mediated Hydroxymethylation in Reprogramming to Induced Pluripotency

By

Tao Wang (王韬)

5-Hydroxymethylcytosine (5hmC) is a newly discovered modified form of cytosine that has been suspected to be an important epigenetic modification in stem cells and during neurodevelopment. Here, we report the roles of 5-hmC during reprogramming as well as neurodevelopment. Mammalian somatic cells can be directly reprogrammed into induced pluripotent stem cells (iPSCs) by introducing defined sets of transcription factors. Somatic cell reprogramming involves epigenomic reconfiguration, conferring iPSCs with characteristics similar to embryonic stem cells (ESCs). Human ES cells contain 5hmC, which is generated through the oxidation of 5-methylcytosine by the TET enzyme family. Here we show that 5hmC levels increase significantly during reprogramming to human iPSCs mainly owing to TET1 activation, and this hydroxymethylation change is critical for optimal epigenetic reprogramming, but does not compromise primed pluripotency. Compared with hES cells, we find iPS cells tend to form large-scale (100 kb-1.3 Mb) aberrant reprogramming hotspots in subtelomeric regions, most of which display incomplete hydroxymethylation on CG sites. Strikingly, these 5hmC aberrant hotspots largely coincide (~80%) with aberrant iPS-ES non-CG methylation regions. Our results suggest that TET1-mediated 5hmC modification could contribute the epigenetic variation of iPSCs and iPSC-hESC differences.

TET-mediated Hydroxymethylation in Reprogramming to Induced Pluripotency

By

Tao Wang (王韬)

B.S., University of Science and Technology of China, Hefei, China 2005

Advisor: Stephen T. Warren, Ph.D.

Peng Jin, Ph.D.

A dissertation submitted to the Faculty of
the James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology

2013

ACKNOWLEDGEMENTS

It has been 5 years since I joined the Genetics and Molecular Biology PhD program at Emory University. First, I want to thank Dr. Xiao-Jiang Li and Dr. Jeremy Boss in the admission committee who gave me an opportunity to study here. I want to thank Dr. Stephen Warren for his financial support throughout my graduate study. I also want to express my sincere gratitude to Dr. Peng Jin, and Dr. Hao Wu. Under their guidance, I was not only able to get projects finished, but also obtain adequate training.

The most important thing I learned from Dr. Stephen Warren is always starting with a significant question that can lead to novel discoveries, which somehow are even more meaningful than solid downstream research. An unexpected discovery is always more important than whatever follow-up studies.

I learned solid experimental biology from Dr. Peng Jin and his lab members (especially, Dr. Yujing Li) when working for him. Besides, there are many personal skills I can learn from him. Peng is a very positive person and has a very good leadership skill. He knows how to identify people who are appropriate to work for, to work with, and to have whom work for him. I can still remember one time when I asked his lab member, "What is the most important quality do you think of Peng as the boss?", he replied, "Fair and generous". This is just one of examples reflects his personality and leadership.

Furthermore, during the time I pursued my secondary Master's degree in Computer Science here, I was trained extensively by Dr. Hao Wu for genomics research; these skills will be a great plus for my biology background in the future. Hao is a very

nice, intelligent, and genuine person. I really appreciate his time for sharing his biostatistics and bioinformatics knowledge to me and guiding my projects.

Next, I would like to thank my rest committee members, Dr. Anthony Chan, Dr. Stephen Dalton and Dr. Victor Corces for their time and suggestions. Also, I would like to thank my collaborators: Dr. Karen Conneely, Dr. David Cutler, Dr. Viren Patel, Ian Goldlust and Dr. Katie Rudd at the Department of Human Genetics, Dr. I-Ping Chen, Dr. Stormy Chamberlain, Dr. Renhe Xu at the UConn Health Science Center, Dr. Qiang Chang at the University of Wisconsin, Dr. Chuan He, Yu Miao, ChunXiao Song at the University of Chicago, Dr. Ranhui Duan in China and many others.

I would also like to thank the members in the Warren Lab for their help: Dr. Steve Bray, Dr. He Gong, Dr. Joshua Suhl, Dr. Mika Kinoshita, Michael Santoro, Ann Dodd, Tamika Malone, Julie Mowrey and Dr. Yaran Wen. I especially want to thank Steve Bray for helping my projects and scientific writing during my early years in the lab. Importantly, I gradually learned from him more about the meaning of human being, the meaning of life, and that a man with a pure soul for the inner peace is unbeatable. Furthermore, I want to thank Dr. Yujing Li, Dr. Keith Szulwach, Jian Li, Dr. Li Lin, Dr. Xuekun Li and Craig Street in the Jin lab for their help.

It is only during your hard times that you know who your true friends are. I thank Fuchang Yin and Haoran Li for their help on my Computer Sciences master degree studies. I cannot get it done without their genuine help. Finally, this Ph.D. work could not have been accomplished without the continual support and the love of my wonderful other half, Shiyao.

DEDICATION

To Shiyao Wang

for her love and support

Table of Contents

Chapter 1: Toward pluripotency by reprogramming: looking back, looking forward

Introduction 1
Abstract 2
Introduction 3
Reprogramming factors 5
Effect of stoichiometry 8
DNA methylation and de-methylation 10
Small molecule-mediated reprogramming 12
iPS and ES cells differences 14
Elite, stochastic, and deterministic models 16
MicroRNAs in somatic reprogramming 17
Disease modeling 19
Concluding remarks 22
References 25

Chapter 2: Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency

Abstract 34
Introduction 35
Results 37
TET1-mediated hydroxymethylation plays a critical role during reprogramming to pluripotency in human cells 37
5hmC epigenomic landscape during reprogramming 38
5hmC is bi-directionally correlated with DNA methylation changes and associated with pluripotency related gene networks 39
Sequence preferences of 5hmC modification during reprogramming 41
Aberrant 5hmC reprogramming hotspots cluster in telomere-proximal regions 41
Concordance of large-scale 5hmC hotspots and iPS-ES non-CG DMRs 43
Base-resolution 5hmC analyses reveal large-scale hotspots are mainly caused by aberrant CG hydroxymethylation 44
Discussion 47
Methods 51
References 99

Chapter 3: Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum

Abstract 106

Introduction 107
Results 110
Dynamics of DNA hydroxymethylation and its genomic features in human cerebellum 110
5-hmC genomic features during cerebellum development 111
The fetus-specific DhMRs displays pluripotent epigenetic memories 112
DhMRs are associated with genes involved in neurodevelopmental disorders 113
Discussion 117
Material and Methods 122
References 141

Chapter 4: Discussion and future directions

Potential distinct roles between TET1 and TET2 147
DNA methylation and demethylation 149
Future direction for hydroxymethylation studies 150
References 153

List of Tables

Table 2-1. Summary of large-scale hotspots between iPSCs and hESCs 91

Table 2-2. Summary of iPSC lines used in this study 92

Table 2-3. Summary of 5hmC sequencing statistics 93

Table 2-4. DhMRs pairwise comparison between fibroblast biological replicates, and between iPSCs and original fibroblasts 94

Table 2-5. Summary of quantitative RT-PCR primers used in this study 95

Table 2-6. Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 10 and corresponding amplicon coordinates 96

Table 2-7. Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 18 and 22, and corresponding amplicon coordinates 97

List of Figures

Figure 1-1. Multiple ways of achieving human pluripotent stages. 23

Figure 1-2. DNA methylation and demethylation during reprogramming. 24

Figure 2-1. TET1 is associated with increased hydroxymethylation during human iPSC reprogramming 64

Figure 2-2. TET1 is associated with increased hydroxymethylation during human iPSC reprogramming, but reduction of TET1 does not compromise the pluripotency of human iPS cells. 66

Figure 2-3. Reprogramming confers a 5hmC epigenome in a pattern with a bias towards telomere proximal regions in autosomes. 68

Figure 2-4. 5hmC is associated with gene activity and pluripotency regulatory networks in stem cells. 70

Figure 2-5. Bimodal distribution of 5hmC around TSS and TES. 72

Figure 2-6. Sequence preference of hydroxymethylation modification during reprogramming. 74

Figure 2-7. Aberrant 5hmC reprogramming hotspots cluster at subtelomeric regions. 77

Figure 2-8. 5hmC DhMRs largely overlap with non-CG-DMRs in a large-scale pattern. 79

Figure 2-9. The 5hmC aberrant reprogramming hotspots are not due to genomic instability. 81

Figure 2-10. Large-scale incomplete hydroxymethylation hotspots are characteristics of human iPS cells. 83

Figure 2-11. Large-scale DhMRs in iPS cells are more variable than in hES cells. 84

Figure 2-12. Correlation and confirmation analyses between TAB-Seq and 5hmC capture approach. 86

Figure 2-13. Large-scale hotspots are caused predominantly by aberrant CpG hydroxymethylation. 88

Figure 2-14. Low level of 5hmC at peri-centromeric non-CG DMRs. 90

Figure 3-1. Global DNA hydroxymethylation dynamics in the developing human cerebellum. 127

Figure 3-2. 5-hmC is strongly associated with gene regions. 128

Figure 3-3. Unique genomic features of dynamic 5-hmC in fetal and adult cerebellums. 129

Figure 3-4. Genomic features of DhMRs during cerebellum development. 130

Figure 3-5. Gene ontology analysis reveals that cerebellum 5-hmC related development dynamics are associated with cellular functional groups. 132

Figure 3-6. Fetus-specific DhMRs show more epigenetic memories that are present in embryonic stem cells. 134

Figure 3-7. FMRP target genes have strong 5-hmC enrichment and are significantly associated with DhMRs. 136

Figure 3-8. SFARI autism candidate genes are preferentially associated with developmentally dynamic hydroxymethylation regions. 138

Figure 3-9. Autism suggestive and strong evidence genes are significantly associated with dynamic 5-hmCs. 140

Figure 4-1. Similarity between non-CG mC and 5hmC modification. 152

Chapter 1

Introduction

Toward pluripotency by reprogramming: looking back, looking forward

This invited review manuscript will be published in ***Protein & Cell***.

Abstract

The somatic epigenome can be reprogrammed to a pluripotent state by a combination of transcription factors. Altering cell fate involves transcription factors cooperation, epigenetic reconfiguration, such as DNA methylation and histone modification, posttranscriptional regulation by microRNAs, and so on. Nevertheless, such reprogramming is inefficient. Evidence suggests that during the early stage of reprogramming, the process is stochastic, but by the late stage, it is deterministic. In addition to conventional reprogramming methods, dozens of small molecules have been identified that can functionally replace reprogramming factors and significantly improve induced pluripotent stem cell (iPSC) reprogramming. Indeed, iPS cells have been created recently using chemical compounds only. iPSCs are thought to display subtle genetic and epigenetic variability; this variability is not random, but occurs at hotspots across the genome. Here we discuss the progress and current perspectives in the field. Research into the reprogramming process today will pave the way for great advances in regenerative medicine in the future.

2. Introduction

In the development of multicellular organisms, a single fertilized cell gives rise to different types of cells with distinct functions. The classic view of cell fate specification is that the undifferentiated, totipotent or pluripotent state is at the top of the multiple types of differentiated somatic states. Conrad Hal Waddington was the first to describe lineage specification in terms of an epigenetic landscape (Goldberg et al., 2007; Waddington, 1957). Metaphorically, a progenitor cell undergoing terminal differentiation is like a marble rolling down a landscape: the marbles will slide downhill, compete for grooves, and eventually come to rest at the lowest points. These lowest points represent the different cell fates. Since marbles tend not to roll back, when cells become more committed during normal development, the cell differentiation potential becomes more restricted. Because Waddington's model fits well in almost all cases, lineage commitment and differentiation has long been considered unidirectional and irreversible.

However, Gurdon showed that the somatic epigenome can be reprogrammed to pluripotency via nuclear reprogramming (Gurdon et al., 2003). Nuclear reprogramming in mammalian cells was first achieved by somatic cell nuclear transfer (SCNT), which established that a nucleus from an adult somatic cell can be reprogrammed by an unfertilized enucleated oocyte(Wilmut et al., 1997). The SCNT experiment was the first evidence that pluripotency can be restored from terminally differentiated cells, and showed that the developmental process is reversible. Subsequently, another form of reprogramming, cell fusion, in which adult somatic cells are fused with ES cells or embryonic germ (EG) cells, was used to reset the somatic epigenome to a pluripotent

state (Cowan et al., 2005; Tada et al., 1997; Tada et al., 2001). These experiments raise an unanswered and interesting question: which gene product(s) in an enucleated oocyte, ES cells or EG cells are the critical factors in reprogramming.

By screening 24 pluripotency factors, in 2006, Takahashi and Yamanaka showed that only four factors, *Oct4*(O), *Sox2*(S), *Klf4*(K), and *c-Myc*(M), when used in combination via retrovirus delivery, can convert somatic fibroblasts to embryonic-like stem cells, or induced pluripotent stem cells(iPSCs) (Figure 1-1) (Takahashi and Yamanaka, 2006). Thereafter, forced expression of different combinations of genes was shown to successfully reprogram fibroblasts, peripheral blood, keratinocytes, and many other types of somatic cells into iPS cells in many species including humans(Aasen et al., 2008; Giorgetti et al., 2009; Haase et al., 2009; Loh et al., 2009; Seki et al., 2010; Staerk et al., 2010; Sun et al., 2009; Takahashi et al., 2007). The delivery methods of these transgenes has expanded as well; among them now are lentivirus, sendai virus, mRNA, episome vectors, and synthetic self-replicative RNA, to name a few (Fusaki et al., 2009; Warren et al., 2010; Wernig et al., 2008; Yoshioka et al., 2013; Yu et al., 2009). Compared with SCNT, the transcription factor-mediated cellular reprogramming process is long, inefficient, and the epigenome variation of iPSCs is large. Many studies have focused extensively on these and illuminated many expected and unexpected mechanisms in this simple scheme, but complicated process. In this review we will summarize the molecular mechanisms of cellular reprogramming, the different methods for efficient reprogramming, and compare iPSC and ESC equivalence.

3. Reprogramming factors

Reprogramming is a dedifferentiation process, which is the reverse of cell differentiation. In normal development, pluripotent cells appear transiently; however, ES cells can self-renew and maintain pluripotency *in vitro*. This suggests ES cells are blocked by particular epigenetic roadblocks. Therefore, during the dedifferentiation process, reprogramming factors push the cells up into the pluripotent state bypassing the epigenetic road blocks. The four factors O, S, K and M must be expressed in correct stoichiometry that provides a sufficient push, as well as in the right direction. Once they reach the pluripotent state, cells must be blocked by an epigenetic barrier so they can remain. In rare situations as represented by inefficient reprogramming, some cells after reprogramming could be blocked by epigenetic barriers and thus acquire self-renew-ability and become capable of differentiating into multiple lineages.

It is thought that OSKM primarily bind their putative binding sites, alter the corresponding gene expression, and change cell fate. Direct evidence for this is that partially reprogrammed cells, which represent an intermediate reprogramming stage, have failed to activate some pluripotency regulators. In these cells, OCT4, SOX2, and KLF4 primary targeting is impaired, and genes that are specifically co-bound by O, S, K lack binding and are transcriptionally silenced (Sridharan et al., 2009). Nevertheless, the mechanism would seem to be more complicated, as reprogramming efficiency increases significantly when cells are infected with highly expressed OSKM. Higher expression of transcription factor is known to increase the strength of nonspecific or low-affinity binding. This phenotype suggests the possibility that low-affinity or random binding sites by OSKM may also play an important role. In tumor cells, elevated c-Myc is found to

bind low-affinity E-box-like sequences, which in turn leads to increased levels of transcription (Lin et al., 2012). Similarly, one could predict that OSKM may also have low-affinity binding sites in ES cells, and the binding may have biological consequences. Yet whether it is stochastic binding or low-affinity binding that is crucial or rate limiting for reprogramming is still unknown.

Among the reprogramming factors OCT4, SOX2 and KLF4, most binding events happen primarily in closed chromatin, which consists of condensed heterochromatin (Soufi et al., 2012). Among those three, OCT4 is found to be the most critical reprogramming factor. OCT4 mainly inhibits the expression of differentiation-related genes in ESCs (Kim et al., 2008; Pardo et al., 2010). When OCT4 is combined with certain chemical compounds, it is sufficient to convert somatic cells into iPSCs. The binding of O, S, K to closed chromatin and the subsequent alteration of it early in reprogramming may therefore be a critical step, because the binding affinity for condensed chromatin for most transcription factors is low, thus they are unable to access the specific sequence. Unlike O, S, K, c-Myc is not essential for reprogramming, but it does increase the efficiency of iPS colony formation. For c-Myc, the binding is biased towards active and open chromatin, which is marked by H3K4 methylation (Soufi et al., 2012). c-Myc is also found to bind to closed chromatin, but this requires O, S, K binding. These data suggest that c-Myc is not a main initiating factor, but rather a positive modulating factor for the other three reprogramming factors.

Activation of endogenous *Oct4* and *Nanog* are crucial for establishing iPSCs. In addition to local regulation such as the alteration of chromatin states by OSKM, DNA looping or non-local interaction also determines the pluripotency of the stem cells. There

are two potential mechanisms. One is that looping affects the expression of key pluripotent genes by promoting enhancer and promoter interaction. For example, there is a cohesin-complex-mediated intrachromosomal loop that links a downstream enhancer to *Oct4*'s promoter, enabling activation of *Oct4* transcription (Zhang et al., 2013). Also, in another study, KLF4 was found to organize long-range chromosomal interactions with the *Oct4* locus, suggesting the reprogramming factors like KLF4 can directly regulate long-range interaction (Wei et al., 2013). The second mechanism is represented by Nanog promoter cis regulation. *Nanog* promoter regions interact with many loci genome-wide and are important for regulating reprogramming via this interaction. A large number of these loci are bound by mediator or cohesin. The establishment of *Nanog* interactions during reprogramming often precedes the transcriptional upregulation of associated genes, suggesting the interaction is important for reprogramming. Depletion of these mediators or cohesin results in a disruption of contacts and the acquisition of a differentiation stage interaction pattern (Apostolou et al., 2013).

In addition to OSKM, pluripotency can also be induced by many combinations of transcriptional factors, such as pluripotency associated factors and maternal factors, including *Nanog*, *Lin28*, *Glis1*, *Esrrb*, *Tbx3* and *Utf1* (Feng et al., 2009; Han et al., 2010; Maekawa et al., 2011; Yu et al., 2007; Zhao et al., 2008). In the case of *Glis1*, it can efficiently generate iPS cells together with OSK. *Glis1* is highly expressed in unfertilized oocytes and one-cell stage embryos. When in combination with OSK, *Glis1* promotes the expression of multiple pro-reprogramming factors, including *Myc*, *Nanog*, *Lin28*, *Wnt*, *Essrb*, and factors involved in the mesenchymal to epithelial transition (Maekawa et al., 2011). Furthermore, the basal transcription machinery, including the transcription factor

IID (TFIID) complex, affects reprogramming efficiency of fibroblasts and is involved in maintaining the pluripotent state. Overexpression of TFIID subunits greatly enhances reprogramming (Pijnappel et al., 2013). All these findings suggest that reprogramming factors need to inhibit lineage specifiers, which are considered to be pluripotency rivals and involved in linear commitment, to convert to pluripotent state. Unexpectedly, a recent study identified eight mesendodermal lineage specifiers as *Oct4* substitutes: *Cebpa*, *Hnf4a*, *Gata3*, *Gata4*, *Gata6*, *Grb2*, *Pax1*, and *Sox7*. *Oct4* and its substitutes attenuated the elevated expression of ectodermal genes, such as *Dlx3*, which were triggered by *Sox2*, *Klf4*, and c-*Myc* (Shu et al., 2013). Their findings present the first evidence that lineage specifiers can replace reprogramming factors as well as facilitate reprogramming. The underlying model is that lineage specifiers, such as *Oct4* replacements, act to balance with other mutually exclusive lineage specifiers such as *Sox2*. As a result, lineage specifiers synergistically influence the induction of pluripotency.

4. Effect of stoichiometry

Interestingly, the four factors stoichiometry-the relative expression level of the four factors- can significantly influence both reprogramming efficiency and the quality of the resulting iPS cells. Higher expression of *Oct4* than the other three factors will generate more iPSC colonies; the reverse ratio will decrease the efficiency (Papapetrou et al., 2009; Tiemann et al., 2011). Moreover, differences in the order of OSKM polycistronic vector can cause a significant quality difference in iPSCs. When expressed polycistronically in the order of OKSM, the expression of *c-Myc* and *Sox2* are found to be higher, and the

Dlk1-Dio3 imprinting locus on mouse chromosome 12qF1 is aberrantly silenced in most of the iPSC clones (Stadtfeld et al., 2010a). Loss of imprinting at the Dlk1-Dio3 locus has been associated with lower pluripotency including poor chimera formation and failure to generate all-iPSC mice by tetraploid complementation. Furthermore, the incidence of tumors in mice created by iPSCs in the order of OKSM is higher (Stadtfeld et al., 2010b). While in the order of OSKM, there is higher expression of *Oct4* and *Klf4* and lower expression of *c-Myc* and *Sox2*, and the reprogrammed iPSCs harbor an active Dlk1-Dio3 locus, which is similar to ESCs. The order of OSKM also produces iPS cells that efficiently generate all-iPSC mice by tetraploid complementation, and do not create mice with tumors (Carey et al., 2011). These studies demonstrate that the stoichiometry of reprogramming factors is critical for epigenetic transformation: a skewed combination will lead to poor-quality iPS cells. Importantly, the sequential introduction of reprogramming factors, such as *Oct4-Klf4* first, then *c-Myc* and finally *Sox2* at the first several days of reprogramming outperforms simultaneous induction (Liu et al., 2013). This suggests that *Oct4* and *Klf4* may have higher expression than *Sox2* and *c-Myc* at the beginning of the reprogramming process, meaning the stoichiometry may primarily have effects in the early stage of reprogramming.

Once pluripotency is established, on the contrary, a reduced *Oct4* expression level seems to enhance pluripotency. Oct4+/- ESCs show increased genome-wide binding of OCT4, particularly at pluripotency-associated enhancers, and increase homogeneous expression of pluripotency transcription factors such as *Nanog* by reducing *Nanog*-low and *Nanog*-negative cells. Thus reduced *Oct4* expression enhances ES or iPS cells self renewal, and delays differentiation (Karwacki-Neisius et al., 2013).

5. DNA methylation and demethylation

The iPSC methylome is different from the somatic methylome (Deng et al., 2009; Lister et al., 2009). In mammals, DNA methylation predominantly occurs at cytosine on CpG sites. In embryonic stem cells, up to 25% of methylation can also occur on non-CpG sites (Laurent et al., 2010; Lister et al., 2009). This is particularly interesting, as it predisposes to the function of non-CpG methylation. Non-CpG methylation tends to occur at exonic regions of actively transcribed regions. The exact function of non-CpG methylation in mammals remained unknown. DNA methylation is catalyzed by DNMT3a/b and maintained by DNMT1(Smith and Meissner, 2013). DNMT3a/b are believed to be *de novo* DNA methyltransferase. DNMT3a/b deficient MEFs can generate iPS cells, and their depletion moderately decreases efficiency compared to wild-type MEFs, suggesting *de novo* methylation during reprogramming is not essential and plays only a minor role (Pawlak and Jaenisch, 2011). Interestingly, *de novo* methylation by DNMT3a and DNMT3b is critical during the developmental process and the reprogramming of germ cells (Kato et al., 2007; Okano et al., 1999).

In contrast, DNA demethylation plays a major role in determining iPS cells transformation processes (Figure 1-2). During reprogramming, the activation of endogenous Oct4, Nanog, and many other pluripotent genes is accompanied by demethylation of cytosines at their promoter or enhancer regions. Insufficient demethylation of these promoter/enhancer regions leads to partially reprogrammed cells. Furthermore, the inhibition of DNA methylation by DNMT1 inhibitors can increase

reprogramming efficiency (Mikkelsen et al., 2008). All this evidence suggests DNA methylation acts as a major barrier to cellular reprogramming, and DNA demethylation plays an important role in successful reprogramming.

There are two proposed mechanisms of DNA demethylation in cells: a DNA replication-independent active DNA demethylation, and a DNA replication-dependent passive DNA demethylation. In the scenario of DNA replication-dependent demethylation, reprogramming factors or some of their targets might antagonize the activity of Dnmt1 or its binding partner, UHRF1, which in turn leads to the progressive loss of DNA methylation with cell division (Bostick et al., 2007; Sharif et al., 2007). The putative DNA active demthylation pathway was found during last several years. In this pathway, Ten-eleven translocation (TET) proteins sequentially catalyze cytosine to 5-hydroxycytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (He et al., 2011; Ito et al., 2011; Iyer et al., 2009; Tahiliani et al., 2009). The 5fC and 5caC may in turn be removed by TDG and replaced by cytosine via base excision repair (BER) pathway enzymes (He et al., 2011; Maiti and Drohat, 2011; Schiesser et al., 2012). This lead to the proposal that TET proteins may function as DNA demethylases. This cycle was found to be feasible at least biochemically *in vitro*. TET proteins have been intensively studied in ES cells. *TET1* is highly expressed in human ES cells, and *Tet1* and *Tet2* are highly expressed in mouse ES cells. It has been shown that TET1/2 depletion will compromise reprogramming efficiency (Costa et al., 2013; Doege et al., 2012; Gao et al., 2013; Wang et al., 2013), although TET1 and TET2 proteins are not required for essential pluripotency, and are dispensable for maintaining ES cells. One possible mechanism is that TET1 and TET2 interact with NANOG, enhancing the demethylation

of Oct4 and other pluripotent gene promoters and enhancers. Specifically, one study showed that *Tet1* could replace *Oct4*, to induce iPSCs (Gao et al., 2013). Interestingly, in human iPSCs, *TET2* is not expressed. Whether TET2 has a unique role during reprogramming in mouse ES cells remains unknown. In another model, the deaminase Aid (or Aicda) was proposed recently to play a role in demethylation. It can demethylate the *NANOG* and *OCT4* promoters after cell fusion of mouse ESCs and human fibroblasts (Bhutani et al., 2010; Popp et al., 2010). Furthermore, Aid, through the regulation of Mbd4 and Gadd45, is involved in DNA demethylation in zebrafish (Rai et al., 2008). However, because Aid expression is low in ESCs and iPSCs, whether it has major role in iPS cell reprogramming is unclear. Recently, Aid was reported to act to remove epigenetic memory, and Aid-null somatic cells fail to stabilize pluripotency in the later stage of the reprogramming process (Kumar et al., 2013). Further research should reveal to what extent active demethylation contributes to overall DNA demethylation.

6. Small molecule mediated reprogramming

Conventional reprogramming methods use viruses or transgenes, which not only pose the risk of future reactivation, but can also cause insertion mutagenesis. As a result, conventional reprogramming methods result in iPS cells that are potentially tumorigenic. This risk of cancer may limit iPSC clinical applications. Furthermore, iPSCs may trigger immune rejections (Zhao et al., 2011). A chemical approach that uses small molecules to generate iPS cells may reduce these safety concerns about them. First, chemical approaches are presumably non-immunogenic. In addition, small molecules can easily

pass through cell membranes, so they can be removed after they have initiated the reprogramming. Using proper compounds like those are FDA approved should minimize the risk of mutation. To date, dozens of small molecules have been identified that can functionally replace reprogramming factors and significantly improve iPSC reprogramming (Huangfu et al., 2008a; Huangfu et al., 2008b; Li et al., 2009; Shi et al., 2008a; Shi et al., 2008b). They primarily target cell signaling pathways, such as the TGFβ pathway, and nuclear epigenetic factors. One example is BIX-01294, a methyltransferase G9a inhibitor, which can replace *Sox2* and *c-Myc* for reprogramming (Shi et al., 2008a; Shi et al., 2008b). A-83-01, a TGFβ receptor inhibitor, enhances MEF reprogramming; in combination with AMI-5, a protein arginine methyltransferase inhibitor, it enables reprogramming of MEFs transduced with *Oct4* only (Yuan et al., 2011).

Many studies have managed to reduce the number of genes needed to reprogram cells by using small-molecule chemical compounds, but those cases always required Oct4. Recently, iPS cells were created using chemical compounds only; these were called chemically induced iPS cells (CiPSC) (Figure 1-1) (Hou et al., 2013). Using a cocktail of seven compounds, this group was able to get 0.2% of cells to convert, an efficiency comparable to those from standard iPS production techniques. Moreover, the chemical factors were able to induce iPSCs from both mouse embryonic fibroblasts and adult fibroblasts. These small molecules include: CHIR, a glycogen synthase kinase 3 inhibitor; 616452, a TGF-beta inhibitor; FSK, a cAMP agonist; DZNep, an S-adenosylhomocysteine hydrolase inhibitor; TTNPB, a synthetic retinoic acid receptor ligand; valproic acid, a histone deacetylase inhibitor; and tranylcypromine (or Parnate),

an inhibitor of lysine-specific demethylase 1. Some of these inhibitors target unexpected pathways, which will reveal other unknown aspects of the reprogramming process. Nevertheless, a detailed comparison of the CiPS and ES cells is needed to determine whether there are subtle differences between them and whether these differences are functionally important for downstream applications.

7. iPS and ES differences

iPSCs are functionally equivalent to ESCs. ESCs and iPSCs share key features of pluripotency, including the expression of pluripotency markers, the ability to differentiate into germ layers, teratoma formation in immunodeficient mice, and tetraploid complementation for mouse iPS cells. The key question is whether there are subtle difference between iPSCs and ESCs, and if so, does this lead to biological consequences. The transcriptomes, proteomes, and epigenomes of ESCs and iPSCs have been compared, and results suggest iPSCs may be different from ESCs, leading to concerns about the differentiation potentials of each individual line and the safety of iPSCs for therapeutic applications (Bock et al., 2011; Chin et al., 2009; Liang and Zhang, 2013; Lister et al., 2011; Nazor et al., 2012; Ruiz et al., 2012; Wang et al., 2013). Here we will explore the issue from an epigenetic perspective. The study results above have lead to three models of the equivalence between iPSCs and ESCs. The first model states that there are small but consistent differences between ESCs and iPSCs (Chin et al., 2009; Stadtfeld et al., 2010a); in this model, the differences are unique to iPSCs or to ESCs, and thus could be used as a marker to distinguish iPSCs from ESCs. As discussed earlier, the Dlk3-Dio

locus was believed to be inactive in mouse iPSCs and was proposed as a marker of iPSCs; however, it turned out the phenotype was caused by a skewed expression level of reprogramming factors. The second model states that iPSCs and ESCs should be treated as two largely overlapping groups that share unique genetic and epigenetic features. In this model, iPSCs show more epigenetic variance, and each iPSC may represent a unique epigenetic status with variable differentiation potential; however, each individual iPSC line cannot be distinguished from ESC lines (Bock et al., 2011; Kim et al., 2011; Lister et al., 2011).Therefore, based on these observations, many believe there are no differences between the iPSC and ESC populations. A third model, and perhaps the more likely one, given new evidence, is that iPSCs display subtle genetic and epigenetic variability. Most importantly, this variability is not random, but only occurs at certain genes or loci, forming aberrant reprogramming hotspots. Not all iPSCs have aberrant events in all these hotspots, but experience events in different combinations of hotspots. For example, hotspot regions with incomplete 5hmC/non-CG methylation tend to cluster in telomere-proximal regions (Wang et al., 2013). Also, in a separate study, gene expression in some iPSCs with aberrant 5hmC in these genes is different than in ESCs (Ruiz et al., 2012). An intriguing finding is that megabase domains of H3K9me3, which impairs OSKM binding and reprogramming, largely overlap with 20 reprogramming hotspots (Soufi et al., 2012). These H3K9me3 domains are refractory to OSKM binding at the initial 24 hours after reprogramming. This suggests a possible mechanism: these reprogramming hotspots are resistant to OSKM binding, fail to recruit histone demethylase, and are subsequently incapable of initiating TET and DNMT3a/b recruitment. There are fewer

aberrant hotspots than megabase domains of H3K9me3, suggesting that malfunction of those aberrant hotspots is less critical for iPS cell survival.

8. **Elite, stochastic, and deterministic models.**

Because iPSC reprogramming efficiency is very low, only a small fraction of cells will transform into iPSCs. After Yamanaka's report, some researchers suspected that only a few somatic cells are competent for reprogramming. In this "elite" model, these rare somatic stem cells were contaminated in donor cells and generated the iPSCs, while the differentiated cells would be resistant to reprogramming. However, several lines of evidence show this is not true. First, subsequent improvements in the methods of reprogramming resulted in efficiencies as high as 10-20%. It is unlikely that tissue stem cells comprise this high a percentage of somatic cells. Secondly, iPSC colonies have been derived from terminally differentiated B and T cells (Hochedlinger and Jaenisch, 2002; Seki et al., 2010). In T cells, specific genomic rearrangement of the immunoglobulin locus or the T cell receptor in iPSC cells proved that the cells were derived from mature B or T cells, but not the mesenchymal stem cells. Lastly, one study indicated that over 90% of terminal differentiated B cells have the potential to generate daughter cells that eventually become iPSCs (Hanna et al., 2009).

Ruling out the elite model, left the question of whether the reprogramming process is stochastic or deterministic. The stochastic model states that somatic cells have to go through the various epigenetic blocks to become iPSCs. In the stochastic model, most differentiated cells have the potential to become iPS cells, however, whether or

when a given cell would become an iPSC cell cannot be predicted. In the deterministic model, reprogrammed cells would be generated with a fixed timescale; SCNT is generally considered to fit the deterministic model. More evidence now supports both models for iPSC reprogramming. At early stage, the reprogramming is stochastic as supported by clonal cell analysis (Hanna et al., 2009). Moreover, single-cell gene expression profiling at various stages demonstrates cells from an early stage become iPSCs with variable latency (Buganim et al., 2012). Although reprogramming is stochastic, early activation of some pluripotent genes, such as *Esrrb*, *Utf1*, *Lin28*, and *Dppa2*, may determine cells to become iPSCs. In somatic cells, many essential pluripotency loci are marked with H3K9me3, such as *Nanog, Dppa4, Sox2, Gdf3,* and *Prdm14* (Polo et al., 2012; Samavarchi-Tehrani et al., 2010; Soufi et al., 2012). These genes are refractory to OSKM binding at early stage and are activated later in reprogramming process. Acquisition of the final pluripotent state requires a later stabilization stage marked by the expression of those pluripotency markers (Golipour et al., 2012). Activation of these H3K9me3 marked loci is crucial for reprogramming to full iPSCs, suggesting that, once activated, the cell transits from a stochastic to a deterministic stage (Chen et al., 2013; Soufi et al., 2012). In summary, evidence suggests that during the early stage, the reprogramming is a stochastic process, and when it reaches the late stage, it is deterministic.

9. MicroRNA in somatic reprogramming

MicroRNAs are a family of small non-coding RNAs that bind to partially complementary sequences in messenger RNAs, inducing mRNA degradation or translational silencing(Bartel, 2009). Changing somatic cell fate to a pluripotent state requires a complete chromatin reorganization to allow the activation of an endogenous program that sustains self-renewal while preventing differentiation. The reprogramming is accompanied by miRNA expression changes. miRNAs have been implicated in the regulation of the self-renewal and differentiation potential of pluripotent stem cells. For example, dgcr8-null mESCs, in which miRNA biogenesis is impaired, have a reduced proliferation rate, and fail to induce differentiation (Wang et al., 2007). Thus, it is not surprising that a subset of miRNAs is required for efficient and essential reprogramming, while others act as reprogramming "roadblocks". MiRNAs required for efficient and essential reprogramming have similar targeting sequences, and may therefore regulate downstream targets cooperatively. Examples include miR-291-3p, -294, -295, and -302d, which increase reprogramming efficiency with *Oct4*, *Klf4* and *Sox2* (Judson et al., 2009). These miRNAs are the ES cell-specific cell cycle regulating micorRNAs, which increase reprogramming by accelerating the G1 to S phase transition during cell cycle (Wang et al., 2008). In contrast, overexpressing “roadblock miRNAs”, like miR-21 and -29a, impede reprogramming (Yang et al., 2011). The p53 and ERK1/2 pathways are regulated by miR-21 and miR-29, which in turn modulate reprogramming.

Interestingly, studies have shown that miRNAs alone, without any exogenous factors, can generate iPS cells, possibly even more effectively than transcription factors (Anokye-Danso et al., 2011; Miyoshi et al., 2011). The first study employed a lentivirus delivery system producing miRNA cluster 302/367. MiR367 expression activates Oct4

gene expression and suppresses Hdac2. Moreover, miR-302-targeted co-suppression of four epigenetic regulators, AOF2 (KDM1/LSD1), AOF1, MECP1-p66 and MECP2, could cause global DNA demethylation (Lin et al., 2011). The second study directly transfected mature miRNAs with a combination of miR-200c, miR-302s and miR-369s family miRNAs. Both approaches successfully produced mouse and human iPS cells from fibroblasts. Nevertheless, there is a discrepancy for miRNA cluster 302/367 in reprogramming. In MEFs by piggybac transfer, microRNA cluster 302/367 could not generate iPSCs (Lu et al., 2012), while another study using human adipose stem cells and failed to produce iPSCs by delivering miRNA-302s alone (Hu et al., 2013). These discrepancies could be caused by different delivering systems. For example, it was found that miR-302-induced reprogramming is dosage dependent (Lin et al., 2011), so the microRNA concentration must be within a specific range.

10. Disease modeling

iPS technology has opened new possibilities for human genetic disease modeling. Before the iPSC era, obtaining human pluripotent stem cells carrying a particular genetic mutation was mired in ethnical issues, because it required isolating ES cells from and the destruction of blastocysts (Revazova et al., 2007). Now, by reprogramming cells from a simple skin biopsy or blood, researchers can generate iPS cells from patients with any disease. iPS technology is not merely a replacement for hESC study, because it overcomes two obstacles associated with hESCs: ethical concerns about the use of human embryos and potential immune rejection after non-autologous therapeutic transplantation.

The possibility of generating pluripotent cells from patient somatic cells and subsequently differentiating them into the desired cell types will give us new insights into the pathogenesis of a broad spectrum of diseases (Merkle and Eggan, 2013). iPS cell lines from patients with different syndromes have been successfully established and differentiated into defective cell types related to disease (Cherry and Daley, 2013; Onder and Daley, 2012). By comparing disease specific iPS cell lines to their healthy or normal counterparts, we can study the biological mechanisms for genetic variants that affect the risk and progression of the disease. Using this approach has yielded novel insights into various diseases with either Mendelian or complex inheritance, among them Alzheimer's disease(Yagi et al., 2011), Parkinson's disease (Hargus et al., 2010; Park et al., 2008; Soldner et al., 2009), amyotrophic lateral sclerosis(ALS) (Dimos et al., 2008; Mitne-Neto et al., 2011), Down syndrome(Li et al., 2012) and schizophrenia(Brennand et al., 2011). The most rigorous way to study the effects of genetic variants in human disease would be the generation of isogenic iPSCs, which differs only in the mutation and has the same genetic background. These disease-specific iPS cells and isogenic control cells would also enable screening for novel drugs (Engle and Puppala, 2013). In addition, human disease cell types derived from iPSCs would be more relevant for toxicological testing during the drug development process, compared with the established cancer origin cell types or animal models used now.

Reprogramming of somatic cells into iPS cells also holds tremendous promise for regenerative medicine, the process of replacing damaged tissue. iPSCs can potentially differentiate into any type of cell, and since they are genetically identical to the patients, presumably will not be immunogenic. This holds out the hope of treating patients who

need regenerative therapies, including disorders characterized by the loss or destruction of cells or tissues, such as the loss of dopaminergic neurons in Parkinson's disease, autoimmune destruction of beta cells in type 1 diabetes, and spinal cord injury, to name a few (Yu et al., 2013). In the case of Parkinson's disease, a degenerative disorder of the central nervous system, patients progressively lose nerve cells that produce dopamine, causing a loss of motor function. In this new avenue of treatment, the aim is to create iPS cells from a patient, differentiate these cells into the dopamine-producing neurons that have been destroyed by disease, and transplant the cells created in the dish back into the patient's brain.

iPSCs will also be valuable for providing patient-specific cellular therapy by generating autologous iPS cells through reprogramming. In this method, gene defects in patient-specific iPSCs would be corrected by methods like ZFN, TALEN, or CRISPR (Gaj et al., 2013), the iPSCs differentiated into the disease-relevant cells, and the cells returned back to the patient. This avenue of therapy will offer the prospect of treatments for a broad range of disorders. For example, using a ZFN technology, researchers reported a sequence of events for successfully correcting a mutation in human iPSCs derived from individuals with α1-antitrypsin deficiency(A1ATD) due to a point mutation (Glu342Lys) in α1-antitrypsin (Yusa et al., 2011). A1ATD is an autosomal recessive disorder that results in liver cirrhosis and represents the most common inherited metabolic disease of the liver. Researchers first took adult skin cells, reprogrammed the adult cells to iPSCs, corrected the gene mutation in both alleles with ZFN, and differentiated the cells *in vitro* into hepatocyte-like cells. They demonstrated that these

corrected hepatocyte-like cells were able to colonize the liver in mouse and had functional activities.

11. Concluding remarks

Taken together, reprogramming by transcriptional factors not only supports the idea that cell fate changes can be bidirectional and reversible, but also opens new opportunities for the study of cell transdifferentiation. Importantly, studying iPSCs has broadened our understanding of cellular differentiation/dedifferentiation mechanisms, also yielding valuable information for disease modeling and clinical applications. The recently created all-chemically induced iPS cells will facilitate this application process. We know iPS cells are not exactly equal to ES cells, and whether the subtle differences are consequential for iPSC clinical applications remains unclear. Recently, researchers achieved the reprogramming of human somatic cells into pluripotent embryonic stem cells by SCNT (Tachibana et al., 2013), making an important step for iPSC study. It will be interesting to see whether stem cells derived from SCNT are more like embryonic stem cells.

12. Acknowledgements

We thank Cheryl Strauss, Michael Santoro, and Helen Gong at the Department of Human Genetics for helpful reading of the manuscript. This study was supported by the National Institutes of Health (NS079625 and HD073162 to P.J.).

Figure 1-1

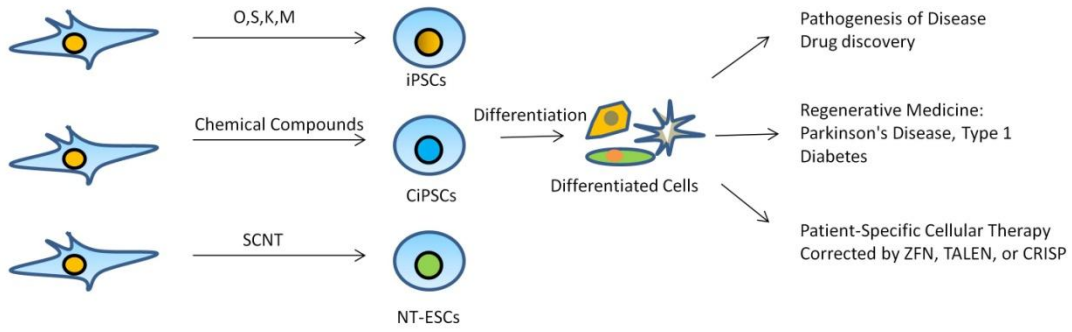

Figure 1-1. Multiple ways of achieving human pluripotent stages. (1). Transcription factors, such as OCT4, SOX2, KLF4, and c-Myc mediated reprogramming; (2). Reprogramming to chemically induced iPS cells by the small-molecule combination VC6TFZ; (3). Reprogramming human somatic cells into pluripotent embryonic stem cells by SCNT. These reprogrammed stem cells have opened new possibilities for human genetic disease modeling, hold tremendous potential for regenerative medicine, and enable patient-specific cellular therapy, by which gene defects in patient-specific iPSCs would be corrected by methods like ZFN, TALEN, or CRISPR.

Figure 1-2

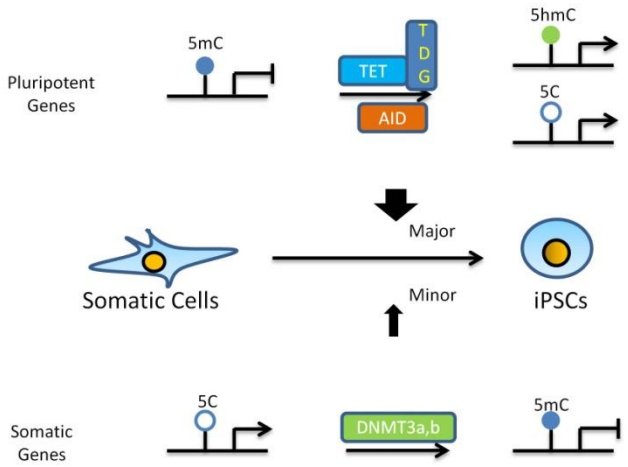

Figure 1-2. DNA methylation and demethylation during reprogramming. *De novo* DNA methylation during reprogramming is not essential and plays only a minor role. Depletion of DNMT3a and 3b moderately decreases reprogramming efficiency compared to wild-type cells. In contrast, DNA demethylation plays a major role and determines iPS transformation processes. TET1/2 depletion compromises reprogramming efficiency. A second possible pathway for demethylation involves the deaminase Aid (or Aicda). *Aid*-null somatic cells fail to stabilize the pluripotency in the later stage during the reprogramming process.

References:

Aasen, T., Raya, A., Barrero, M.J., Garreta, E., Consiglio, A., Gonzalez, F., Vassena, R., Bilic, J., Pekarik, V., Tiscornia, G., *et al.* (2008). Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nature biotechnology 26, 1276-1284.

Anokye-Danso, F., Trivedi, C.M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., Zhang, Y., Yang, W., Gruber, P.J., Epstein, J.A., *et al.* (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell stem cell 8, 376-388.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., *et al.* (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell 12, 699-712.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell 136, 215-233.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature 463, 1042-1047.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., Ziller, M., Croft, G.F., Amoroso, M.W., Oakley, D.H., *et al.* (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 144, 439-452.

Bostick, M., Kim, J.K., Esteve, P.O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 317, 1760-1764.

Brennand, K.J., Simone, A., Jou, J., Gelboin-Burkhart, C., Tran, N., Sangar, S., Li, Y., Mu, Y., Chen, G., Yu, D., *et al.* (2011). Modelling schizophrenia using human induced pluripotent stem cells. Nature 473, 221-225.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 150, 1209-1222.

Carey, B.W., Markoulaki, S., Hanna, J.H., Faddah, D.A., Buganim, Y., Kim, J., Ganz, K., Steine, E.J., Cassady, J.P., Creyghton, M.P., *et al.* (2011). Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. Cell Stem Cell 9, 588-598.

Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Wu, Y., Guo, L., *et al.* (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nat Genet 45, 34-42.

Cherry, A.B., and Daley, G.Q. (2013). Reprogrammed cells for disease modeling and regenerative medicine. Annual review of medicine 64, 277-290.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., *et al.* (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5, 111-123.

Costa, Y., Ding, J., Theunissen, T.W., Faiola, F., Hore, T.A., Shliaha, P.V., Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S., *et al.* (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature 495, 370-374.

Cowan, C.A., Atienza, J., Melton, D.A., and Eggan, K. (2005). Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. Science 309, 1369-1373.

Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E.M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.H., Yu, J., *et al.* (2009). Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 27, 353-360.

Dimos, J.T., Rodolfa, K.T., Niakan, K.K., Weisenthal, L.M., Mitsumoto, H., Chung, W., Croft, G.F., Saphier, G., Leibel, R., Goland, R., *et al.* (2008). Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science 321, 1218-1221.

Doege, C.A., Inoue, K., Yamashita, T., Rhee, D.B., Travis, S., Fujita, R., Guarnieri, P., Bhagat, G., Vanti, W.B., Shih, A., *et al.* (2012). Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. Nature 488, 652-655.

Engle, S.J., and Puppala, D. (2013). Integrating human pluripotent stem cells into drug development. Cell Stem Cell 12, 669-677.

Feng, B., Jiang, J., Kraus, P., Ng, J.H., Heng, J.C., Chan, Y.S., Yaw, L.P., Zhang, W., Loh, Y.H., Han, J., *et al.* (2009). Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nat Cell Biol 11, 197-203.

Fusaki, N., Ban, H., Nishiyama, A., Saeki, K., and Hasegawa, M. (2009). Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome. Proc Jpn Acad Ser B Phys Biol Sci 85, 348-362.

Gaj, T., Gersbach, C.A., and Barbas, C.F., 3rd (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31, 397-405.

Gao, Y., Chen, J., Li, K., Wu, T., Huang, B., Liu, W., Kou, X., Zhang, Y., Huang, H., Jiang, Y., *et al.* (2013). Replacement of Oct4 by Tet1 during iPSC induction reveals an important role of DNA methylation and hydroxymethylation in reprogramming. Cell Stem Cell 12, 453-469.

Giorgetti, A., Montserrat, N., Aasen, T., Gonzalez, F., Rodriguez-Piza, I., Vassena, R., Raya, A., Boue, S., Barrero, M.J., Corbella, B.A., *et al.* (2009). Generation of induced pluripotent stem cells from human cord blood using OCT4 and SOX2. Cell stem cell 5, 353-357.

Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007). Epigenetics: a landscape takes shape. Cell 128, 635-638.

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C.L., Trcka, D., and Wrana, J.L. (2012). A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. Cell Stem Cell 11, 769-782.

Gurdon, J.B., Byrne, J.A., and Simonsson, S. (2003). Nuclear reprogramming and stem cell creation. Proceedings of the National Academy of Sciences of the United States of America 100 Suppl 1, 11819-11822.

Haase, A., Olmer, R., Schwanke, K., Wunderlich, S., Merkert, S., Hess, C., Zweigerdt, R., Gruh, I., Meyer, J., Wagner, S., *et al.* (2009). Generation of induced pluripotent stem cells from human cord blood. Cell stem cell 5, 434-441.

Han, J., Yuan, P., Yang, H., Zhang, J., Soh, B.S., Li, P., Lim, S.L., Cao, S., Tay, J., Orlov, Y.L., *et al.* (2010). Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature 463, 1096-1100.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595-601.

Hargus, G., Cooper, O., Deleidi, M., Levy, A., Lee, K., Marlow, E., Yow, A., Soldner, F., Hockemeyer, D., Hallett, P.J., *et al.* (2010). Differentiated Parkinson patient-derived induced pluripotent stem cells grow in the adult rodent brain and reduce motor asymmetry in Parkinsonian rats. Proceedings of the National Academy of Sciences of the United States of America 107, 15921-15926.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., *et al.* (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333, 1303-1307.

Hochedlinger, K., and Jaenisch, R. (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. Nature 415, 1035-1038.

Hou, P., Li, Y., Zhang, X., Liu, C., Guan, J., Li, H., Zhao, T., Ye, J., Yang, W., Liu, K., *et al.* (2013). Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. Science.

Hu, S., Wilson, K.D., Ghosh, Z., Han, L., Wang, Y., Lan, F., Ransohoff, K.J., Burridge, P., and Wu, J.C. (2013). MicroRNA-302 increases reprogramming efficiency via repression of NR2F2. Stem Cells 31, 259-268.

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A.E., and Melton, D.A. (2008a). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol 26, 795-797.

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., and Melton, D.A. (2008b). Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. Nat Biotechnol 26, 1269-1275.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333, 1300-1303.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle 8, 1698-1710.

Judson, R.L., Babiarz, J.E., Venere, M., and Blelloch, R. (2009). Embryonic stem cell-specific microRNAs promote induced pluripotency. Nature biotechnology 27, 459-461.

Karwacki-Neisius, V., Goke, J., Osorno, R., Halbritter, F., Ng, J.H., Weisse, A.Y., Wong, F.C., Gagliardi, A., Mullin, N.P., Festuccia, N., *et al.* (2013). Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell 12, 531-545.

Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Okano, M., Li, E., Nozaki, M., and Sasaki, H. (2007). Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum Mol Genet 16, 2272-2280.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132, 1049-1061.

Kim, K., Zhao, R., Doi, A., Ng, K., Unternaehrer, J., Cahan, P., Huo, H., Loh, Y.H., Aryee, M.J., Lensch, M.W., *et al.* (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol 29, 1117-1119.

Kumar, R., DiMenna, L., Schrode, N., Liu, T.C., Franck, P., Munoz-Descalzo, S., Hadjantonakis, A.K., Zarrin, A.A., Chaudhuri, J., Elemento, O., *et al.* (2013). AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes. Nature 500, 89-92.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., *et al.* (2010). Dynamic changes in the human methylome during differentiation. Genome Res 20, 320-331.

Li, L.B., Chang, K.H., Wang, P.R., Hirata, R.K., Papayannopoulou, T., and Russell, D.W. (2012). Trisomy correction in Down syndrome induced pluripotent stem cells. Cell stem cell 11, 615-619.

Li, W., Zhou, H., Abujarour, R., Zhu, S., Young Joo, J., Lin, T., Hao, E., Scholer, H.R., Hayek, A., and Ding, S. (2009). Generation of human-induced pluripotent stem cells in the absence of exogenous Sox2. Stem Cells 27, 2992-3000.

Liang, G., and Zhang, Y. (2013). Genetic and Epigenetic Variations in iPSCs: Potential Causes and Implications for Application. Cell Stem Cell 13, 149-159.

Lin, C.Y., Loven, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional amplification in tumor cells with elevated c-Myc. Cell 151, 56-67.

Lin, S.L., Chang, D.C., Lin, C.H., Ying, S.Y., Leu, D., and Wu, D.T. (2011). Regulation of somatic cell reprogramming through inducible mir-302 expression. Nucleic Acids Res 39, 1054-1065.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.

Lister, R., Pelizzola, M., Kida, Y.S., Hawkins, R.D., Nery, J.R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., *et al.* (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471, 68-73.

Liu, X., Sun, H., Qi, J., Wang, L., He, S., Liu, J., Feng, C., Chen, C., Li, W., Guo, Y., *et al.* (2013). Sequential introduction of reprogramming factors reveals a time-sensitive requirement for individual factors and a sequential EMT-MET mechanism for optimal reprogramming. Nat Cell Biol 15, 829-838.

Loh, Y.H., Agarwal, S., Park, I.H., Urbach, A., Huo, H., Heffner, G.C., Kim, K., Miller, J.D., Ng, K., and Daley, G.Q. (2009). Generation of induced pluripotent stem cells from human blood. Blood 113, 5476-5479.

Lu, D., Davis, M.P., Abreu-Goodger, C., Wang, W., Campos, L.S., Siede, J., Vigorito, E., Skarnes, W.C., Dunham, I., Enright, A.J., *et al.* (2012). MiR-25 regulates Wwp2 and Fbxw7 and promotes reprogramming of mouse fibroblast cells to iPSCs. PloS one 7, e40938.

Maekawa, M., Yamaguchi, K., Nakamura, T., Shibukawa, R., Kodanaka, I., Ichisaka, T., Kawamura, Y., Mochizuki, H., Goshima, N., and Yamanaka, S. (2011). Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. Nature 474, 225-229.

Maiti, A., and Drohat, A.C. (2011). 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.
Merkle, F.T., and Eggan, K. (2013). Modeling human disease with pluripotent stem cells: from genome association to function. Cell Stem Cell 12, 656-668.
Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49-55.
Mitne-Neto, M., Machado-Costa, M., Marchetto, M.C., Bengtson, M.H., Joazeiro, C.A., Tsuda, H., Bellen, H.J., Silva, H.C., Oliveira, A.S., Lazar, M., *et al.* (2011). Downregulation of VAPB expression in motor neurons derived from induced pluripotent stem cells of ALS8 patients. Human molecular genetics 20, 3642-3652.
Miyoshi, N., Ishii, H., Nagano, H., Haraguchi, N., Dewi, D.L., Kano, Y., Nishikawa, S., Tanemura, M., Mimori, K., Tanaka, F., *et al.* (2011). Reprogramming of mouse and human cells to pluripotency using mature microRNAs. Cell stem cell 8, 633-638.
Nazor, K.L., Altun, G., Lynch, C., Tran, H., Harness, J.V., Slavin, I., Garitaonandia, I., Muller, F.J., Wang, Y.C., Boscolo, F.S., *et al.* (2012). Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell 10, 620-634.
Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247-257.
Onder, T.T., and Daley, G.Q. (2012). New lessons learned from disease modeling with induced pluripotent stem cells. Current opinion in genetics & development 22, 500-508.
Papapetrou, E.P., Tomishima, M.J., Chambers, S.M., Mica, Y., Reed, E., Menon, J., Tabar, V., Mo, Q., Studer, L., and Sadelain, M. (2009). Stoichiometric and temporal requirements of Oct4, Sox2, Klf4, and c-Myc expression for efficient human iPSC induction and differentiation. Proc Natl Acad Sci U S A 106, 12759-12764.
Pardo, M., Lang, B., Yu, L., Prosser, H., Bradley, A., Babu, M.M., and Choudhary, J. (2010). An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. Cell Stem Cell 6, 382-395.
Park, I.H., Arora, N., Huo, H., Maherali, N., Ahfeldt, T., Shimamura, A., Lensch, M.W., Cowan, C., Hochedlinger, K., and Daley, G.Q. (2008). Disease-specific induced pluripotent stem cells. Cell 134, 877-886.
Pawlak, M., and Jaenisch, R. (2011). De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state. Genes Dev 25, 1035-1040.
Pijnappel, W.W., Esch, D., Baltissen, M.P., Wu, G., Mischerikow, N., Bergsma, A.J., van der Wal, E., Han, D.W., Bruch, H., Moritz, S., *et al.* (2013). A central role for TFIID in the pluripotent transcription circuitry. Nature 495, 516-519.
Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., *et al.* (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell 151, 1617-1632.
Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature 463, 1101-1105.

Rai, K., Huggins, I.J., James, S.R., Karpf, A.R., Jones, D.A., and Cairns, B.R. (2008). DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. Cell 135, 1201-1212.

Revazova, E.S., Turovets, N.A., Kochetkova, O.D., Kindarova, L.B., Kuzmichev, L.N., Janus, J.D., and Pryzhkova, M.V. (2007). Patient-specific stem cell lines derived from human parthenogenetic blastocysts. Cloning and stem cells 9, 432-449.

Ruiz, S., Diep, D., Gore, A., Panopoulos, A.D., Montserrat, N., Plongthongkum, N., Kumar, S., Fung, H.L., Giorgetti, A., Bilic, J., *et al.* (2012). Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. Proc Natl Acad Sci U S A 109, 16196-16201.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell 7, 64-77.

Schiesser, S., Hackner, B., Pfaffeneder, T., Muller, M., Hagemeier, C., Truss, M., and Carell, T. (2012). Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. Angew Chem Int Ed Engl 51, 6516-6520.

Seki, T., Yuasa, S., Oda, M., Egashira, T., Yae, K., Kusumoto, D., Nakata, H., Tohyama, S., Hashimoto, H., Kodaira, M., *et al.* (2010). Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. Cell Stem Cell 7, 11-14.

Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., *et al.* (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature 450, 908-912.

Shi, Y., Desponts, C., Do, J.T., Hahm, H.S., Scholer, H.R., and Ding, S. (2008a). Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. Cell Stem Cell 3, 568-574.

Shi, Y., Do, J.T., Desponts, C., Hahm, H.S., Scholer, H.R., and Ding, S. (2008b). A combined chemical and genetic approach for the generation of induced pluripotent stem cells. Cell Stem Cell 2, 525-528.

Shu, J., Wu, C., Wu, Y., Li, Z., Shao, S., Zhao, W., Tang, X., Yang, H., Shen, L., Zuo, X., *et al.* (2013). Induction of pluripotency in mouse somatic cells with lineage specifiers. Cell 153, 963-975.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat Rev Genet 14, 204-220.

Soldner, F., Hockemeyer, D., Beard, C., Gao, Q., Bell, G.W., Cook, E.G., Hargus, G., Blak, A., Cooper, O., Mitalipova, M., *et al.* (2009). Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. Cell 136, 964-977.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 151, 994-1004.

Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. Cell 136, 364-377.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010a). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature 465, 175-181.

Stadtfeld, M., Maherali, N., Borkent, M., and Hochedlinger, K. (2010b). A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nat Methods 7, 53-55.

Staerk, J., Dawlaty, M.M., Gao, Q., Maetzel, D., Hanna, J., Sommer, C.A., Mostoslavsky, G., and Jaenisch, R. (2010). Reprogramming of human peripheral blood cells to induced pluripotent stem cells. Cell stem cell 7, 20-24.

Sun, N., Panetta, N.J., Gupta, D.M., Wilson, K.D., Lee, A., Jia, F., Hu, S., Cherry, A.M., Robbins, R.C., Longaker, M.T., *et al.* (2009). Feeder-free derivation of induced pluripotent stem cells from adult human adipose stem cells. Proceedings of the National Academy of Sciences of the United States of America 106, 15720-15725.

Tachibana, M., Amato, P., Sparman, M., Gutierrez, N.M., Tippner-Hedges, R., Ma, H., Kang, E., Fulati, A., Lee, H.S., Sritanaudomchai, H., *et al.* (2013). Human embryonic stem cells derived by somatic cell nuclear transfer. Cell 153, 1228-1238.

Tada, M., Tada, T., Lefebvre, L., Barton, S.C., and Surani, M.A. (1997). Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells. The EMBO journal 16, 6510-6520.

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., and Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Current biology : CB 11, 1553-1558.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., *et al.* (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930-935.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131, 861-872.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663-676.

Tiemann, U., Sgodda, M., Warlich, E., Ballmaier, M., Scholer, H.R., Schambach, A., and Cantz, T. (2011). Optimal reprogramming factor stoichiometry increases colony numbers and affects molecular characteristics of murine induced pluripotent stem cells. Cytometry A 79, 426-435.

Waddington, C.H. (1957). The strategy of the genes; a discussion of some aspects of theoretical biology (London,, Allen & Unwin).

Wang, T., Wu, H., Li, Y., Szulwach, K.E., Lin, L., Li, X., Chen, I.P., Goldlust, I.S., Chamberlain, S.J., Dodd, A., *et al.* (2013). Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency. Nat Cell Biol 15, 700-711.

Wang, Y., Baskerville, S., Shenoy, A., Babiarz, J.E., Baehner, L., and Blelloch, R. (2008). Embryonic stem cell-specific microRNAs regulate the G1-S transition and promote rapid proliferation. Nature genetics 40, 1478-1483.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet 39, 380-385.
Warren, L., Manos, P.D., Ahfeldt, T., Loh, Y.H., Li, H., Lau, F., Ebina, W., Mandal, P.K., Smith, Z.D., Meissner, A., *et al.* (2010). Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. Cell Stem Cell 7, 618-630.
Wei, Z., Gao, F., Kim, S., Yang, H., Lyu, J., An, W., Wang, K., and Lu, W. (2013). Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. Cell Stem Cell 13, 36-47.
Wernig, M., Lengner, C.J., Hanna, J., Lodato, M.A., Steine, E., Foreman, R., Staerk, J., Markoulaki, S., and Jaenisch, R. (2008). A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol 26, 916-924.
Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., and Campbell, K.H. (1997). Viable offspring derived from fetal and adult mammalian cells. Nature 385, 810-813.
Yagi, T., Ito, D., Okada, Y., Akamatsu, W., Nihei, Y., Yoshizaki, T., Yamanaka, S., Okano, H., and Suzuki, N. (2011). Modeling familial Alzheimer's disease with induced pluripotent stem cells. Human molecular genetics 20, 4530-4539.
Yang, C.S., Li, Z., and Rana, T.M. (2011). microRNAs modulate iPS cell generation. RNA 17, 1451-1460.
Yoshioka, N., Gros, E., Li, H.R., Kumar, S., Deacon, D.C., Maron, C., Muotri, A.R., Chi, N.C., Fu, X.D., Yu, B.D., *et al.* (2013). Efficient Generation of Human iPSCs by a Synthetic Self-Replicative RNA. Cell Stem Cell 13, 246-254.
Yu, D.X., Marchetto, M.C., and Gage, F.H. (2013). Therapeutic translation of iPSCs for treating neurological disease. Cell Stem Cell 12, 678-688.
Yu, J., Hu, K., Smuga-Otto, K., Tian, S., Stewart, R., Slukvin, II, and Thomson, J.A. (2009). Human induced pluripotent stem cells free of vector and transgene sequences. Science 324, 797-801.
Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., *et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science 318, 1917-1920.
Yuan, X., Wan, H., Zhao, X., Zhu, S., Zhou, Q., and Ding, S. (2011). Brief report: combined chemical treatment enables Oct4-induced reprogramming from mouse embryonic fibroblasts. Stem Cells 29, 549-553.
Yusa, K., Rashid, S.T., Strick-Marchand, H., Varela, I., Liu, P.Q., Paschon, D.E., Miranda, E., Ordonez, A., Hannan, N.R., Rouhani, F.J., *et al.* (2011). Targeted gene correction of alpha1-antitrypsin deficiency in induced pluripotent stem cells. Nature 478, 391-394.
Zhang, H., Jiao, W., Sun, L., Fan, J., Chen, M., Wang, H., Xu, X., Shen, A., Li, T., Niu, B., *et al.* (2013). Intrachromosomal Looping Is Required for Activation of Endogenous Pluripotency Genes during Reprogramming. Cell Stem Cell 13, 30-35.
Zhao, T., Zhang, Z.N., Rong, Z., and Xu, Y. (2011). Immunogenicity of induced pluripotent stem cells. Nature 474, 212-215.
Zhao, Y., Yin, X., Qin, H., Zhu, F., Liu, H., Yang, W., Zhang, Q., Xiang, C., Hou, P., Song, Z., *et al.* (2008). Two supporting factors greatly improve the efficiency of human iPSC generation. Cell Stem Cell 3, 475-479.

Chapter 2

Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency

This manuscript has been published on ***Nature Cell Biology***. Tao Wang, Hao Wu, Yujing Li, Keith E. Szulwach, Li Lin, Xuekun Li, I-Ping Chen, Ian S. Goldlust, Stormy J. Chamberlain, Gene Ananiev, Ann Dodd, He Gong, Ji Woong Han, Youngsup Yoon, M. Katharine Rudd, Chun-Xiao Song, Chuan He, Qiang Chang, Stephen T. Warren*, Peng Jin*. Nature Cell Biology, 2013, doi:10.1038/ncb2748

ABSTRACT

Mammalian somatic cells can be directly reprogrammed into induced pluripotent stem cells (iPSCs) by introducing defined sets of transcription factors. Somatic cell reprogramming involves epigenomic reconfiguration, conferring iPSCs with characteristics similar to embryonic stem cells (ESCs). Human ES cells contain 5-hydroxymethylcytosine (5hmC), which is generated through the oxidation of 5-methylcytosine by the TET enzyme family. Here we show that 5hmC levels increase significantly during reprogramming to human iPSCs mainly due to TET1 activation, and this hydroxymethylation change is critical for optimal epigenetic reprogramming, but does not compromise primed pluripotency. Compared with hES cells, we find iPS cells tend to form large-scale (100 kb-1.3 Mb) aberrant reprogramming hotspots in subtelomeric regions, most of which display incomplete hydroxymethylation on CG sites. Strikingly, these 5hmC aberrant hotspots largely coincide (~80%) with aberrant iPS-ES non-CG methylation regions. Our results suggest that TET1-mediated 5hmC modification could contribute the epigenetic variation of iPSCs and iPSC-hESC differences.

INTRODUCTION

Pluripotency is defined as a stem cell state with the potential to differentiate into any of the three germ layers. Somatic cells can be reprogrammed to a pluripotent state by defined factors such as OCT4, SOX2, KLF4, c-MYC, NANOG and LIN28[1-3]. These iPSCs are extremely similar to ESCs. During the reprogramming process, the global epigenetic landscape in somatic cells has to be reset to reach a pluripotent state via DNA methylation/demethylation and chromatin remodelling processes.

Besides 5-methylcytosine (5mC), which is known to display dynamic changes during early embryonic and germ cell development as well as the reprogramming process, the mammalian genome also contains 5hmC, which is generated by oxidation of 5mC by the TET family of enzymes[4, 5]. The Tet proteins function in ESCs regulation, myelopoiesis and zygote development[6-10]. 5hmC was found to be widespread in many tissues and cell types at different levels[11, 12]. Particularly, 5hmC is abundant in the central nervous system and ESCs. Several reports have explored the genome-wide distribution of 5hmC modification in mES cells and hES cells, and suggest that it is enriched in gene bodies and enhancers[13, 14].

Reprogramming toward pluripotency involves a dynamic epigenetic modification process. 5hmC has been implicated in the DNA demethylation process[15], pointing to a potential role for 5hmC modification during reprogramming toward pluripotency. Thus, understanding the dynamic 5hmC changes during reprogramming will provide additional insight into somatic cell reprogramming mechanisms.

Multiple studies suggest there are subtle yet substantial genetic and epigenetic differences between iPS cells and hES cells[16, 17]. The current consensus is that iPS cells and ES cells are two overlapping classes of heterogeneous cells, with iPS cells being more variable than hES cells[18]. Although iPS cells and hES cells are functionally equivalent in general, the subtle genetic and epigenetic differences could lead to functional consequences among individual lines. Previous study of the base-resolution methylomes of iPSCs and ESCs identified differentially methylated regions (DMRs) between iPSCs and ESCs, consisting of CG-DMRs and non-CG-DMRs[16, 17]. However, the traditional bisulfite sequencing technique they used could not distinguish 5mC from 5hmC[19], which means how these DMRs are caused by hydroxymethylation differences remains unknown.

Here we show that 5hmC levels increase significantly during reprogramming to human iPSCs mainly due to TET1 activation, and this hydroxymethylation change is critical for optimal epigenetic reprogramming. We found that during reprogramming extensive genome-wide 5hmC modification occurs. Importantly, we identified specific aberrant reprogramming hotspots in iPS cells, which cluster on a large-scale (100kb-1.3Mb) at subtelomeric regions bearing incomplete CG hydroxymethylation. These hotspots largely overlap with aberrant non-CG methylation hotspots, suggesting hydroxymethylation contributes to the epigenetic difference between iPS cells and hES cells.

RESULTS

TET1-mediated hydroxymethylation plays a critical role during reprogramming to pluripotency in human cells

DNA methylation is a major barrier to iPS cell reprogramming. Several lines of evidence suggest that 5hmC is involved in the process of DNA demethylation[20, 21]. We found a significant increase of 5hmC level in human iPS cells compared to their original fibroblasts, with the amount in iPSCs being similar to hES cells (Fig. 2-1a).

TET family proteins (TET1, TET2 and TET3) could convert 5mC to 5hmC[6]. We found a statistically significant increase of TET1 and TET3; with a more dramatically increase of TET1, and a slight decrease of TET2 expression (Fig. 2-1b). RNA-seq reveals that TET1 is at a comparable level to NANOG in pluripotent cells, but the expression of TET2 and TET3 are significantly lower (Fig. 2-1c). Depletion of TET1 but not TET2 and TET3 by siRNA could significantly decrease total 5hmC levels in human iPS cells (Fig. 2-1d and Fig. 2-2a,b). Therefore, we conclude that TET1 is the main TET protein regulating hydroxymethylation during human iPS cells reprogramming.

Because cellular reprogramming is an epigenetic state reconfiguring process, we next asked whether TET1-mediated hydroxymethylation changes are critical in human iPSC reprogramming. Introducing shTET1 lentivirus with "Yamanka factors" infection could decrease alkaline phosphatase positive colonies when compared with equal titer shControl lentivirus transduction (Fig. 2-1e,f and Fig. 2-2c,d). shTET1 treated colonies during reprogramming can be further stably maintained, showing decreased TET1 levels,

but similar pluripotent gene expression levels compared with iPSCs (Fig. 2-1g). Furthermore, iPS cells depleted with TET1 maintained a normal undifferentiated stem cell morphology, are positive for alkaline phosphatase, expressed same level pluripotent related factors and stained positive for the pluripotency markers such as NANOG, SOX2, TRA-1-81 (Fig. 2-1h and Fig. 2-2e-g). Therefore, TET1-mediated hydroxymethylation modification is required for optimal induction of iPSCs, but does not compromise the essential pluripotency of human stem cells.

5hmC epigenomic landscape during reprogramming

We employed 5hmC Capture-Seq to assess genome-wide 5-hmC distributions during reprogramming[11]. The cell lines and sequencing statistics are summarized on Table 2-2 and 2-3. Pearson correlation and cluster analysis of the global 5hmC modification pattern suggests a significant difference between iPS cells and fibroblasts (Fig. 2-3a and Table 2-4).

Based on a negative binomial model for testing differential expression of sequencing data[22], we found 267,664 regions in the genome showing differential 5-hydroxymethylation modification between iPS cells and fibroblast (false discovery rate (FDR): 0.01), which denoted as differential 5-hydroxymethylated regions (DhMRs). Among them, 231,866 are hyperDhMRs (5hmC level is higher in iPS cells), and 35,798 are hypoDhMRs (5hmC level is lower in iPS cells) (Fig. 2-3b). The hyperDhMRs show higher gain of 5hmC than the loss of 5hmC observed at hypoDhMRs (Fig. 2-3c). The hyperDhMRs are distributed across all autosomes, but largely missing in sex

chromosomes (Fig. 2-3d). Particularly, of the top 20000 hyperDhMRs (ranked by adjusted p-values), they have a higher probability ($p<0.0001$) of being located in the telomere proximal regions (Fig. 2-3e), as shown by example of Chromosome 1 and Chromosome X (Fig. 2-3f).

5hmC is bi-directionally correlated with DNA methylation changes and associated with pluripotency related gene networks

The analysis described above suggests a global hydroxymethylation change during reprogramming. 5hmC has been suggested linked with gene expression in ES cells and neurons[13, 14, 23-26]. To assess the correlation between 5hmC modifications and gene expression changes during reprogramming, we stratified genes into 9 categories based on gene expression changes between iPS cells and fibroblasts (category 1: high expression in iPS cells, low expression in fibroblast; category 2: medium expression in iPS cells, low expression in fibroblast, *etc*). We then quantified the amount of 5hmC around transcription start site (TSS). As a result, those 9 categories can be clustered into 3 distinct patterns (Fig. 2-4a). Of note, most expressed genes during reprogramming show a bimodal distribution with a depletion of 5hmC in TSS sites, whereas genes remain silenced after reprogramming show a peak in TSS sites. Among 3 clusters, cluster1 has the lowest 5hmC levels in TSS; cluster 3 has the highest levels of 5hmC in TSS, but has lowest 5hmC levels in gene bodies (Fig. 2-4b).

We then examined the correlation between absolute amount of transcripts and 5hmC enrichment. We noticed that hyperDhMRs tend to form bimodal distribution associated

with gene activity in iPS cells, with the lowest level similar to the level in fibroblast in TSS regions (Fig. 2-4c and Fig. 2-5). TES regions also show a bimodal distribution, the depletion is more dramatic in a narrower region centred on TES (Fig. 2-5). However, compared with hypoDhMRs, hyperDhMRs are more enriched in TSS, exons and TES (Fig. 2-6a). We observed a significant negative correlation between 5hmC level of TSS surrounding regions (±200bp) and gene expression levels in iPS cells (Fig. 2-6b).

We also observe bidirectional correlation between 5hmC level and DNA methylation during reprogramming process. 80% of the partially methylated domains (PMD), which displays lower levels of CG methylation in somatic cells than stem cells[27], have increased 5hmC levels, with the rest have no 5hmC level change (Fig. 2-4d). Interestingly, we also found around 60% stem cells hypoDMRs (lower CG methylation in stem cells) shows increased 5hmC modification (Fig. 2-4b). Collectively, our results suggest that increased hydroxymethylation not only occur in loci with increased methylation but also loci with decreased methylation during reprogramming.

Based on the results of bimodal distribution of 5hmC in TSS and TES, we then determined whether this distribution is associated with core pluripotency regulatory networks. We found that pluripotent master regulators, such as OCT3/4 and NANOG, bear this typical modification in iPSCs but not in fibroblasts (Fig. 2-4e). We further tested the relation of 5hmC and key pluripotency factors binding sites[27]. We found a more than 8-fold higher than expected overlap between 5hmC-enriched regions and OCT4, KLF4 binding sties, with a weak association with NANOG and SOX2 binding

sites (Fig. 2-4f). Our results suggest that OCT4 and KLF4 regulatory networks may require 5hmC to regulate pluripotency during reprogramming. Furthermore, gene ontology analysis shows that genes acquiring most 5hmC are involved in stem cell differentiation and patterning process (Fig. 2-4g), suggesting 5hmC in stem cells are highly correlated with pluripotency.

Sequence preferences of 5hmC modification during reprogramming

We compared the CG, CH (CA, CT, CC), CHG preference of hyperDhMRs and hypoDhMRs. HyperDhMRs tend to be located at higher C and G enriched regions, as well as CHG and CH enriched regions, whereas hypoDhMRs have the same level as the genome background (Fig. 2-4h). Previous observations suggest that 5hmC modification is related to CpG-density[24, 28]. We find that in iPSCs, the low CpG content group of CpG islands tend to have more 5hmC modifications (Fig. 2-6c), which is consistent with the observation that DNA methylation occurs more frequently in CpG islands with low CpG content[29]. Furthermore, 5hmC modifications acquired during reprogramming tend to occur within the unique sequence in which the methylation is evolutionarily less conserved[30](Fig. 2-6d-f).

Aberrant 5hmC reprogramming hotspots cluster in telomere-proximal regions

Reprogramming of somatic cells to a pluripotent state requires complete reversion of the somatic epigenome into the pluripotent epigenome, which is an ES-like-state. iPSCs retain some type of somatic memory from their previous identity[31-33]. We further determined the genome-wide 5hmC modification differences between iPS and ES cells,

aiming to understand whether 5hmC modifications underlie the differences between hES cells and iPS cells. To reduce the biases of tissue origins, we used 9 iPS cells derived from different origins, 6 of which are from fibroblasts as mentioned earlier, 2 are derived from peripheral blood cells, and 1 is derived from human exfoliated deciduous teeth cells (SHED).

In general, global DNA hydroxymethylation patterns are very similar between iPS and ES cells (Fig. 2-7a). A comprehensive analysis of 372,423 5hmC-enriched regions between 4 hES cell and 9 iPS cell lines led to the identification of 113 iPS-ES-DhMRs that were differentially hydroxymethylated in at least one iPS cell or ES cell line (FDR<0.01), as shown for the SIGLEC6 and SIGLEC 12 locus in Fig. 2-8a. Surprisingly, these regions are not randomly located across the genome; instead, they tend to cluster at the telomere-proximal regions, in particular, at chromosome 3, 7, 8, 12, and 20 (Fig. 2-7b).

In contrast to the symmetric pattern of DMRs between iPS and ES cells[17], 105 of the 113 iPS-ES DhMRs are hypo-hydroxymethylated, with 5hmC levels similar to their respective progenitors blood cells or fibroblast (Fig. 2-7c,d). Of these DhMRs, the 5hmC patterns are more variable compared with hES cells (Fig. 2-7d). Unsupervised hierarchical clustering using the top 1,000 most variable 5hmC modified regions among all samples could not distinguish hESCs from hiPSCs, suggesting that the variability among iPSCs is not due to different levels of pluripotency, and the 5hmC deviation of iPSCs is not a key determinant to distinguish hESCs from iPSCs (Fig. 2-7e).

Copy number variation (CNV) has been reported to contribute to the variations of iPSCs[34,35]. Since DhMRs cluster at subtelomeric regions and shows depletion of hydroxymethylation, we further examined whether the DhMRs were simply due to genetic variation, such as CNV, instead of real aberrant 5hmC epigenetic modification. To this end we used high-density comparative genomic hybridization (aCGH) array to examine 3 iPSCs and 2 human ESCs. Array CGH yields an average of 70 CNVs on autosomes, none of which is overlapping with the iPS-ES-DhMRs we identified (Fig. 2-9). Therefore, iPS-ES-DhMRs are caused by aberrant epigenetic modification.

Concordance of large-scale 5hmC hotspots and iPS-ES non-CG DMRs

Our results suggest that iPS-ES-DhMRs tend to cluster at telomere proximal regions, forming aberrant reprogramming hotspots. To better define these large-scale regions, we developed a statistical method to identify potential large-scale aberrant reprogramming hotspots. An aberrant reprogramming hotspot is defined as a genomic region satisfying the following conditions: (1) large variability of 5hmC levels among iPS cells, (2) the average 5hmC difference between iPSCs and ESCs is statistically significant, and (3) longer than100kb. 20 large scale regions were identified. Among them, 19 are hypoDhMRs, all of which have the same epigenetic status as their parent cells, pointing to a "somatic memory" during reprogramming, and 1 is hyperDhMRs (Table 1).

We then compared DhMRs with the DMRs identified previously using whole-genome single base bisulfite sequencing, which would not be able to distinguish 5mC from

5hmC[17]. Of the total 113 DhMRs, only 5 overlap with 1,175 CG-DMRs (Fig. 2-8b). Surprisingly, out of the 19 hypo large-scale hotspots, 84.2% overlap with the 24 mega-scale hypo-non-CG-DMRs, whereas the expected percentage is 1.6% based on permutation (Fig. 2-8c). Fig. 2-8d shows one of these regions, chr10: 132010002-133270002, 5-mCH are depleted in iPS cells but not hESC lines; similarly, of the 9 total iPS cells, only iPS-S1 and iPS-S2 derived from blood bear similar levels of 5hmC compared with hESC counterparts. Of note, the variances from iPS cells are significantly larger than ES cells (Fig. 2-10a and Fig. 2-11a, b). None of the iPS cell lines has all of the 19 hypo large-scale DhMRs restored the same level as the 4 human ES cell lines (Fig. 2-10b). This indicates that these large-scale regions tend to form aberrant reprogramming hotspots that were resistant to reprogramming. We did not observe a statistically significant ($p=0.54$) correlation between passage number of iPSCs and the number of aberrant hotspots (Fig. 2-11c), implying that passage number may not be a key determinant of hotspots number in each iPSC line.

The aberrant 5hmC reprogramming hotspots we identified may also explain the transcription level variability in iPSCs. Notably, some of the genes such as TCERG1L and FAM19A (Table 1), have been reported to be expressed at a significantly lower level in many but not all iPSCs as compared to ES cells[36, 37].

Base-resolution 5hmC analyses reveal large-scale hotspots are mainly caused by aberrant CG hydroxymethylation

The observed extremely high concordance between hypo large-scale DhMRs and non-CG-DMRs is surprising, and might indicate that of the previously identified aberrant 5mCH hotspot regions, a significant portion of CH consists of 5hmC; alternatively, these regions could contain both non-CG (mC) and CG (hmC) aberrant modification. The majority of 5hmC in ESCs is found at CG sites[38]. In addition, 5hmC quantification by Tet-Asisted-Bisulfite sequencing (TAB-Seq) and the chemical capture approach is well correlated both genome-widely and within the 20 large-scale hotspots regions (Fig. 2-12a,b). Therefore, it is very likely that the aberrant 5hmC is caused by CG modification.

To test this possibility experimentally, we applied TAB-Seq, which can detect hydroxymethylation status at base resolution, to 2 hESCs and 4 iPS cell lines. We performed base-resolution analysis of 5hmC in 3 randomly chosen large-scale regions, chr10, chr18, chr22, and amplified 5hmC enriched regions by PCR (Fig. 2-13a and Table 2-6,7). We then subjected them to deep sequencing. Deep sequencing of PCR amplicons after traditional bisulfite conversion confirmed that there is epigenetic variation in non-CG sites but not CG sites (Fig. 2-13b,d). Consistent with the results obtained by capture method, we saw the similar 5hmC variations in iPS cells (Fig. 2-13c and Fig. 2-12c,d). Importantly, this incomplete hydroxymethylation is caused by CG modification, but not CH modification (Fig. 2-13c and Fig. 2-12c,d). For example, in the Chr10 hotspot, iPS-B22 and B23 show incomplete 5hmC in CG dinucleotides, but not in CH dinucleotides (Fig. 2-13e). Therefore, our results suggest the coexistence of aberrant non-CG methylation and CG aberrant hydroxymethylation in subtelomeric hotspots (Fig. 2-13f). The concordance of aberrant CG hydroxymethylation with those aberrant CH large-scale

regions suggests there might be crosstalk between epigenetic pathway regulates hydroxymethylation and pathway regulates CH methylation; this crosstalk may behave more stochastically in those subtelomeric regions.

DISCUSSION

Our study suggests that the significant increase of 5hmC during reprogramming is mainly due to the activation of TET1 protein in human iPS cells, which is in contrast to the previous observations that both Tet1 and Tet2 are upregulated in mouse iPS cells. Mouse ESCs are different from human ESCs in many aspects, such as X-chromosome inactivation status in female lines[39]. From a cell signaling perspective, human pluripotency (primed pluripotency) depends mainly on FGF and Activin-Nodal signaling pathways, whereas mouse pluripotency (naïve/ground state pluripotency) is maintained by LIF-STAT pathways. The difference between human and mouse TET family proteins involved in reprogramming may be caused by FGF signaling selection of a subpopulation of hiPSCs. Several studies of generating naïve human iPSCs under LIF signaling have been reported[40, 41]. So it is possible that TET1 and TET2 have distinct roles in regulating pluripotency, with TET2 being involved in naïve pluripotency and TET1 functioning in primed pluripotency. On the other hand, it is possible that TET1-mediated 5hmC modification is unique in human regardless of different pluripotent stages. Since TET1/2 is dispensable for maintaining stem cells pluripotency, and their loss are compatible with embryonic and postnatal development[42], it is likely that TET2 expression is not under positive section for stem cell functions during evolution, thus eventually silenced in human pluripotent stages.

Reprogramming induces a remarkable epigenomic reconfiguration throughout the somatic cell genome. Recently, it was shown that TET1 and TET2, in synergy with NANOG, enhance the efficiency of mouse iPS cells reprogramming[43]. Here we show TET1-mediated hydroxymethylation change is critical for optimal human iPS cells

reprogramming. We further show that TET1-mediated-5hmC modification only affects reprogramming efficiency, but does not alter the essential pluripotency in human stem cells. The pathways involving TET1 regulation largely remain unknown. It would be interesting to know whether the known epigenetic factors such as DOT1L, Kdm2b, *etc* [44, 45] which are negative and positive modulators for reprogramming are linked to TET1-regulated hydroxymethylation modification.

Human iPS cells hold great promise for regenerative medicine and for establishing models of specific diseases. iPS and ES cells are known to share key features of pluripotency, including the expression of pluripotency markers, teratoma formation, cell morphology, the ability to differentiate into germ layers, and tetraploid complementation[46]. Two models depict the equivalence, or lack thereof, between iPSCs and ESCs. One model posits there may be small but consistent differences between ESCs and iPSCs, as suggested before[36, 47]; the other model states that iPSCs and ESCs should be treated as two partially overlapping groups that share unique features. In this second model, single iPS cell lines cannot be distinguished from ES cell lines, though iPSCs shows more epigenetic variance. Mounting evidence supports the latter model[16, 17, 32]. Therefore, each iPSC may represent a unique epigenetic status with variable differentiation potential. The cause and degree of variation remain to be determined. Our study integrates the 5hmC epigenomic mark into the investigation of ES-iPS equivalence. We find that 5hmC occurs extensively in iPS cells at levels similar to ES cells, and there are no consistent 5hmC markers that can distinguish iPSCs from hESCs; however, we identified 20 regions in iPSCs that tend to form large scale (100kb-1.3Mb) aberrant

reprogramming hotspots, supporting the current consensus that iPSCs are more epigenetically variable than ESCs. Remarkably, these regions with 5hmC variations tend to cluster in telomere-proximal regions. The close proximity of the hotspots to telomeres indicates there may be a distinct cellular process that could impede the reprogramming process.

Almost none of the DhMRs overlap with CG-DMRs, suggesting CG-DMRs identified previously are primarily caused by DNA methylation. DNA methylation in non-CG contexts is abundant in pluripotent stem cells (mCHG and mCHH, where H = A, C or T), comprising almost 25% of all cytosines at which DNA methylation is identified. Strikingly, ~80% of large-scale iPS-ES DhMR regions coincide with previously reported non-CG DNA methylation aberrant hotspots[17]. Reciprocally, ~50% of non-CG DMRs overlaps with our identified DhMRs. It was reported that non-CG DMRs also occur in the peri-centromeric zones. Notably, these peri-centromeric regions contain low level of 5hmC (stem cells have similar levels of 5hmC as fibroblasts), suggesting cells do not need to establish 5hmC in these regions during reprogramming (Fig. 2-14). Thus, the concordance occurs mainly at telomere proximal regions. By applying TAB-Seq, we show that incomplete hydroxymethylation occur predominantly at CG sites, but not CH sites, suggesting the co-existence of aberrant non-CG methylation and aberrant CG hydroxymethylation in these regions. During reprogramming, both CH methylation and hydroxymethylation need to be established *de novo* from the somatic epigenome. It is known that non-CG cytosine methylation is exclusively catalysed by Dnmt3a and Dnmt3b[48]. The concordance suggests there might be crosstalk between epigenetics

pathways that regulate the activities of TET and DNMT3, which may behave more stochastically in those subtelomeric regions.

In summary, our results indicate that TET1-mediated 5hmC modification contributes to both the human iPS cell reprogramming process and differences between iPSCs and hESCs. In particular, we identified 20 large-scale aberrant hotspots, suggesting iPSCs are more epigenetically variable than ESCs in terms of 5hmC modification. Our data suggest that, when studying aberrant epigenetic reprogramming events, as well as their functional consequences, at the DNA level, 5hmC modification merits particular consideration, in addition to 5mC.

METHODS

iPSC Reprogramming and Cell Culturing

Human fibroblasts IMR90 and CRL2097 were obtained from ATCC, and GM0011 was obtained from Coriell Cell Repositories. The fibroblasts were cultured in DMEM medium containing 10% FBS, 1× Non-Essential amino acids, 1× glutamine, and 1× Pen/Strep. The H1 hESC and iPSC-IMR90 were obtained from WiCell, Wisconsin. HUES48, HUES49 and HUES53 were obtained from the Human Embryonic Stem Cell Collection at Harvard University. The cells were maintained in hESC/hiPSC standard medium (DMEM/F12, 20% KnockOut Serum Replacement, 1× MEM Non-Essential Amino Acids, 1× glutamine, 0.11 mM 2-mercaptoethanol, 10 ng/ml bFGF) on irradiated MEF feeders.

We focus on efficient reprogramming methods mainly by retrovirus, lentivirus and Sendai virus, all known to have distinct behaviours in establishing iPS cells[49]. Since the stoichiometry of reprogramming factors can influence the epigenetic status of iPS cells[50,51], we included the iPS cells reprogrammed by "Yamanaka factors" and "Thomson factors" either in polycistronic vectors or separate vectors.

For human iPSC-A2, B22, and B23 reprogramming, 2×10^5 fibroblasts were seeded in a well of a 6 well plate on day 1. On day 2, 10ul of concentrated pMXs-hOCT4, hSOX2, c-hMYC and hKLF4 retrovirus were added to cells in the presence of 6 μg/ml Polybrene. A second round of transduction was repeated on day 3. On day 7, the cells were reseeded in 10cm dishes with irradiated MEF feeders. The potential hiPSC colonies were picked between days 18-25. The established iPSC cell lines were subsequently confirmed with

AP staining, pluripotent markers by immunofluorescence staining and the ability to differentiate into 3 germ layers. IPSC-AG2.3 and iPSC-RX35i were reprogrammed in a similar way except: iPSCAG2.3 was derived from fibroblasts transduced with a mixture of hOCT4, hSOX2, hNANOG, and hLIN28 lentiviruses, and iPSC_RX35i were derived from fibroblasts by STEMCCA lentivirus. HiPSCS1 and hiPSCS2 lines were generated from peripheral blood mononuclear cells (PBMNs) of 2 healthy volunteers using Sendai virus (CytoTune-iPS kit; kindly provided by and property of DNAVEC Corp., Japan), which are presumably free of transgene integration. To transduce cells, 4 separate Sendai viruses containing hOCT3/4, hSOX2, hKLF4, and c-MYC were used. Transduced cells were immediately plated onto a 12-well plate. Medium was replaced on day 1; on day 3 cells were trypsinized and passed onto 2 10-cm gelatin-coated culture dishes with irradiated MEFs. Cells were subsequently maintained in iPS medium. iPS colonies were manually isolated based on morphology between day 14 and 30 post infection. HiPSCS3 was obtained by reprogramming stem cells from human exfoliated deciduous teeth (SHEDs) using a STEMCCA lentiviral vector (a generous gift from Dr. Gustavo Mostoslavsky)[52]. Briefly, pulp tissue from a primary upper central incisor was removed and digested in a solution of 1 mg/ml Collagenase/Dispase for 30 min at 37°C. SHED cultures were maintained with alpha-MEM supplemented with 15% FBS, 2 mM glutamine, 100 U/ml penicillin and 100 μg/ml streptomycin until confluent. $5x10^4$ cells were infected with hSTEMCCA-loxP lentivirus for 24 h. Medium was then switched to iPS medium and changed daily for 4 days. Cells were subcultured onto 10-cm gelatin coated culture dishes seeded with irradiated MEFs. iPS colonies were manually isolated between day 20 to 30 post infection.

RNA Interference Experiments

siRNAs targeting TET1, TET2 and TET3 were designed and validated by Dharmacons. siTET1 sequence are: GAUAGGAGAUUAACAUUGG; GCUCAAACGAGGUCCAUUA; ACGAUUAGCUCCAAUUUAU; GACUCUAAUUGGUGUACAA. The iPSCs were dissociated into single cell suspensions by 0.5% trypsin-EDTA. Then 3×10^5 cells were plated on 6-well plates pre-coated with matrigel in mTeSR1 medium with the presence of thiazovivin to increase single cell survival rate. After 24 h, iPSCs were transfected with siRNA by RNAiMax according to the manufacturer's protocol. The final concentration for siRNA is 50 nM or 100nM. 48 h post transfection, the cells were evaluated by qRT-PCR or dot-blot.

To assess TET1 function during reprogramming, CRL2097 fibroblasts were seeded at $1x10^5$ cells per well of a 6 well plate. Cells were transduced with concentrated retrovirus containing Yamanaka factors in two consecutive days. On the first round of infections, cells were infected with equal titer lentivirus either expressing shTET1 or shGFP. Seven days later, cells were reseeded on puromycin-resistant MEF feeders in a 10cm dish in hESC culture medium with puromycin (0.5 μg/ml). Potential iPSC colonies were stained by alkaline phosphatase (Millipore) around 20 days after initial Yamanaka factors infection. pLKO.1-shGFP (control) and pLKO.1-shTET1 lentivirus were made according to standard procedure. Titer of concentrated virus was then determined by QuickTiter Lentivirus Quantitation Kit (Cell Biolabs).

shTET1-75024(Sigma) sequence: CCGGCCCAGAAGATTTAGAATTGATCTCGAGATCAATTCTAAATCTTCTGGGTTTTTG,

pLKO.1-shTET1-75026(Sigma) sequence: CCGGGCAGCTAATGAAGGTCCAGAACTCGAGTTCTGGACCTTCATTAGCTGCTTTTTG. pLKO.1-puro eGFP shRNA(Sigma) was used as control.

5hmC Dot-Blot

DNA was spotted on an Amersham Hybond-N+ membrane (GE Healthcare) and then fixed to the membrane by drying at 80° C for 30 min. The membrane was then blocked with 5% BSA and incubated with polyclonal antibody against 5hmC (1:5000 dilution, Active Motif) as the primary antibody overnight at 4 ℃. Horseradish peroxidase–conjugated secondary antibody against rabbit (1:5000 dilutions, Sigma) was used to incubate the membrane for 1 h at room temperature.

Immunofluorescence Staining

Human iPS cells treated either with shTET1 or shControl were plated onto coverslips that were pre-coated with matrigel in mTeSR1 medium under puromycin selection (0.5 μg/ml). Cells were fixed in 4% paraformaldehyde for 10 minutes. Then cells were permeabilized for 10 minutes with 0.25% Triton X-100. Cells were then incubated for 1 hour with 4% donkey serum blocking buffer. Then cells were incubated with primary antibody over night at 4℃. After washing with PBS with 3 times for 5 minutes, secondary antibodies conjugated to Alexa Fluor-555 (Invitrogen) were used. The primary

antibodies used were Anti-NANOG (Cell Signalling, 3580s, 1:100), SOX2 (Santa Cruz, sc-20088, 1:100), TRA-1-81 (Millipore, 90233, 1:100)

Genomic DNA Preparation and 5hmC Capture

Prior to isolation of genomic DNA, hiPSCs/hESCs were treated with collagenase to detach from feeder cells, and transferred to Matrigel-coated culture plates in mTeSR1 medium (Stemcell Technologies) for at least 3 passages to eliminate the contamination of feeder cells. Genomic DNA was extracted and purified with the DNeasy Kit (Qiagen). Genomic DNA (20 μg-30 μg) was sonicated to an average size of 200bp by the Covaris sonicator. 5hmC labelling reactions were performed according to the previous protocol[11] with some modifications. Briefly, the UDPG-N3 transfer was carried out with 1X reaction buffer containing 50 mM HEPES (pH 7.9) and 25 mM MgCl2, 100 μM of UDP-6-N3-Glu, and 2 μM of wild-type β-glucose transferase for 1 h at 37° C. The labelled DNA was purified by the QIAquick PCR purification kit (Qiagen). Click chemistry was performed with the addition of 150 μM of disulfide-biotin linker, and the mixture was incubated for 2 h at 37° C. The DNA samples were then purified by the Pierce Monomeric Avidin Kit (Thermo) following the manufacturer's recommendations. Subsequently, the 5hmC enriched DNA was concentrated by 10 K Amicon Ultra-0.5 mL Centrifugal Filters (Millipore), then purified and eluted with 12ul H_2O by the MinElute PCR Purification Kit (Qiagen).

.

Quantitative RT-PCR

Total RNA was extracted using the RNeasy Kit (Qiagen). Total RNA (2 μg) was converted to cDNA by using the iScript cDNA Synthesis Kit (Bio-Rad). The cDNA was then diluted by 1:200 and 8ul of each diluted template were subjected to PCR amplification in a 20 μl volume mixed with Power SYBR Green Master Mix (Applied Biosystems). The PCR conditions were an initial 95 ℃ denaturation for 10 min followed by amplification cycles consisting of 95 ℃ for 10 s, 60 ℃ for 60 s, and 72 ℃ for 60 s for 40 cycles. For data analysis, the results were normalized with GAPDH signal. For iPS cells colony analyses, cells were lysed and subjected to reverse transcription by using a Cells-to-Ct kit (Life Technologies) according to the manufacturer's protocol. Primers sequences are listed in Table 2-5.

Library Preparation and Illumina Sequencing for 5hmC Captured DNA

Approximately 50ng of each 5hmC enriched DNA was used for Illumina SR library preparation by NEBNext ChIP-Seq Sample Prep Reagent Set 1(NEB) according to the manufacturer's standard protocol. The sequences of the adapters used for ligation are: 5' P GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG and 5'ACACTCTTTCCCTACACGACGCTCTTCCGATCT and the PCR primers used for the amplification step are: PCR1.1: 5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; PCR2.1: 5' CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT. The concentration of each DNA library was determined in DNA 1000 chip by Agilent 2100 Bioanalyzer. The library concentration was further confirmed with real time PCR by using the DNA Library Quantification Kit (Kapa Biosystems). The libraries were then

diluted to 10nm for sequencing on Illumina HiSeq/ HiScanSQ systems. Briefly, DNA libraries were denatured with NaOH according to the Illumina protocol (final template concentration is 1nM). Denatured libraries were diluted to a final concentration of 6pM. Each sample was then spiked with 1% PhiX control. The libraries were clustered to a single read flow cell according to the Illumina cBot Cluster Generation System procedures. Upon completion of cluster generation the flow cell was run for 50 cycles (36 cycles for H1).

Sequence Alignment and 5hmC Peak Calling

All FASTQ sequence files were aligned to the human reference genome (hg18) using Bowtie 0.12.7[53] with the same criteria: unique genomic matches and no more than 3 mismatches. The aligned tags were further processed to filter the duplicate reads. 5hmC peaks were identified using MACS[54] with the following parameters: effective genome size = 2.7e+09; Tag size= 38 or 50; Bandwidth = 250; P-value cut-off = 1.00e-05 with H1 genomic DNA input as a control.

DhMR Identification

To compare peaks between samples, the 5hmC enriched regions in each sample were coalesced into a union window. We recounted the total aligned reads for each window, then further normalized with each aligned total count. To call differential 5hmC enrichment regions, the Bioconductor Deseq package was used for analysis, and a FDR of 0.01 was used for positive calling. When using iPSCs compared with original fibroblasts, we found significant number of peaks, while using fibroblast compared with a

repeat experiment of fibroblasts only yield background level of peaks, suggesting we captured bona fide 5hmc modification in fibroblasts (Table 2-6).

The large-scale aberrant reprogramming hotspots are defined as genomic regions satisfying the following conditions: (1) the 5hmC levels are highly variable among iPSCs but relatively consistent among ES cells, (2) the average difference of 5hmC levels between iPSCs and ES cells is large, and (3) longer than 100kb. To assess the variability of 5hmC levels, the whole genome was binned into 1kb windows and the read counts within each window were obtained. The biological variation in each window was then calculated using a method of moment estimator (Detailed in Estimating 5hmC Variations). We then smoothed the estimated variance by moving window average with 100kb. The smoothed variances from iPSCs are significantly larger than from ESCs (Fig. 2-11a). We then pooled the smoothed variations of iPSCs and ESCs, and used the 99^{th} quantile as the threshold to detect variable 5hmC regions (VhMR) in iPSCs and ESCs. Thirty-three and one VhMRs are detected from iPSCs and ESCs respectively. We then assessed the average 5hmC levels in these VhMRs. First the counts are normalized by total reads and average 5hmC levels are computed for iPSC and ESC. The average 5hmC levels are greater in ESCs for most VhMRs (Figure 2-11b). The large-scale aberrant reprogramming hotspots are identified as genomic regions satisfying the following criteria: (1) smoothed variances of iPSCs greater than the 99^{th} quantile of the pooled variations; (2) smoothed variances of ESC smaller than the 50^{th} quantile of the pooled variances; (3) differences of smoothed averages between iPSCs and ESCs greater than

the 95th quantile of all absolute differences; and (4) minimum length greater than 100kb. Detected regions closer than 50kb are merged into one.

Estimating 5hmC Variations

We first obtained read counts in non-overlapping 1kb windows. We denote the count for window *i* and sample *j* by X_{ij}. X_{ij} is assumed to follow a negative binomial distribution: $X_{ij} \sim NB(s_j\theta_{ij}, \phi_j)$. Here θ_{ij} is the true 5hmc level, s_j is the library size, and ϕ_j is the dispersion parameter. The negative binomial is a gamma-Poisson compound distribution. It assumes that the true 5hmc level θ_{ij} follows a gamma distribution, and conditional on θ_{ij} the observed counts follow a Poisson distribution. A negative binomial distribution accounts for over-dispersions (sample variance greater than sample mean) so it is often used for modeling sequencing data from biological replicates. The dispersion parameter ϕ_j is the squared coefficient of variation (CV) of the true 5hmc level θ_{ij}, and represents the variability among biological replicates. It can be shown that directly using the sample variances of normalized reads to estimate ϕ_j will lead to erroneous results. Samples with larger library sizes will have smaller variance estimation. We designed the following moment estimator. First, define a new variable $Y_{ij} = (X_{ij}^2 - X_{ij})/S_j$. We have $E[Y_{ij}] = \mu_{ij}^2(\phi_j - 1)$; here μ_{ij} is the expected value of θ_{ij}. We first estimate μ_{ij} as $\hat{\mu}_{ij} = \overline{X_{ij}/S_j}$, then use the method of moment to obtain the estimates for ϕ_j as $\overline{Y_{ij}}/\hat{\mu}_{ij}^2 - 1$. Detailed proofs and derivations for the estimators are presented in a statistical paper[55].

Tet Assisted Bisulfite Based PCR Amplicon Sequencing

To investigate 5hmC distribution at a single-base resolution, hESCs/ iPSCs genomic DNA were subjected to glycosylation and catalyzation by Tet as described previously[38] and the processed DNA were eluted in a ~50ul(500ng) volume. The treated gDNA was bisulfite converted and eluted in 30ul H2O. 1 μl of converted DNA was PCR amplified by using PfuTurbo Cx Hotstart DNA polymerase under the following condition: 2.5U polymerase, 5 μl 10X PfuTurbo Cx reaction buffer, 4 μl 2.5 mM dNTPs, 1 μl primers, The PCR cycling conditions were: 95 ℃ 2 min, 40 cycles of 95 ℃ 30 s, 55 ℃ 30 s, 72 ℃ 1 min, followed by 72 ℃ 5 min. Primers used for amplifying bisulfite converted genomic DNA were designed by Methyl Primer Express® Software v1.0(Invitrogen) targeting chr10, chr18 and chr22 large scale hotspots, and confirmed by specific bands in agarose gel electrophoresis. The average amplicon size is around 200bp. The primer sequence and amplified region coordinates are listed in Table 2-6,7.

The PCR amplicon were further purified by AMPure XP bead, and eluted in 50ul H_2O. The concentration were quantified with a Qubit High Sensitivity kit and then pooled together in equal molar for each sample. Then the mixed amplicons were subjected to library preparation and MiSeq deep sequencing. Briefly, samples were treated by end repair, A-tailing and the ligation of TruSeq adaptors containing compatible indexes by using NEBNext library preparation for Illumina kit. Libraries were then quantified by KAPA SYBR FAST qPCR Kits and pooled together with an equal molar concentration. MiSeq sequencing was performed as standard procedures recommended by Illumina: the concentration of pooled library used was 8pM, and the run was initiated for 2 × 150 bases

of SBS sequencing. Image analysis and base calling were performed with the standard Illumina pipeline. To call mC/5hmC status, the Bismark application was used.

Array CGH

2 µg of HUES48, HUES49, hiPS-B22, hiPS-B23 and hiPS-RX35i DNA were co-hybridized with 2 µg H1 hESC reference DNA to 1x1M Agilent SurePrint G3 Human Catalog oligonucleotide arrays (Agilent Technologies, Santa Clara, CA). The arrays span the entire genome with an oligonucleotide backbone spaced, on average, every 3 kb; the unique identifier (AMADID) for the design is 021529. Arrays were hybridized according to the manufacturer's instructions and scanned using the Agilent high-resolution C scanner (Agilent Technologies). Signal intensities were evaluated using Feature Extraction Version 9.5.1.1 software (Agilent Technologies) and analysed with Genomic Workbench 5.0 software (Agilent Technologies). To detect the maximum number of CNVs, we used a minimum absolute log ratio of 0.25 on at least 4 aberrant probes. To generate figure 2-9, the stringency was raised to 20 aberrant probes.

Genomic Analysis

Microarray data on fibroblasts and iPSCs were obtained from a previous study[37]. The microarray data were normalized and analyzed using Bioconductor's oligo[56] and siggenes packages within R (http://www.r-project.org/). The differentially expressed genes were called by SAM (Significance Analysis of Microarrays) with corrected P-value <0.01.

RefSeq genes and CpG islands were defined based on NCBI build 36/hg18 coordinates downloaded from the UCSC Genome Browser website. Core promoters were arbitrarily

defined as 200bp upstream and downstream of the transcriptional start site of RefSeq genes. Gene bodies are defined as the transcribed regions, from the start to the end of transcription sites for each RefSeq gene.

Association of DhMRs with genomic features was performed by overlapping defined sets of DhMRs with known genomic features obtained from UCSC Tables for NCBI36/hg18: RefSeq Genes, 5′ UTR, Exon, Intron, 3′ UTR, ±500 bp of TSS, RefSeq Intergenic, ±500 bp of TES, CpG Islands. BGC CpG Islands, hypodeaminated CpG islands were defined from a previous study[30]. Transcription factor binding sites for KLF4, NANOG, OCT4, SOX2 and RNA-Seq RPKM values in H1 hESCs were described previously[27]. Data analysis was performed by R (http://www.r-project.org/) scripts.

Genomic views of 5hmC relative enrichment intensity were generated using IGV 2.0.10 and igvtools (Integrated Genomics Viewer tools and browser, http://www.broadinstitute.org/igv/)[57].

ACCESSION NUMBER

Sequencing data have been deposited to GEO with accession number GSE37050.

ACKNOWLEDGEMENTS

We thank Joshua Suhl, Michael Santoro, Steve Bray and Cheryl Strauss for critical reading of the manuscript. We thank Xinping Huang from the Viral Vector Core of the Emory Neuroscience NINDS Core Facilities for preparing the retrovirus/lentivirus used in this study. We are grateful to Julie Mowrey, Viren Patel, Craig Street and Sandeep Namburi for support on Illumina Hiseq2000/Miseq sequencing. This study was supported in part by the National Institutes of Health (NS079625 and HD073162 to P.J.; MH089606 and HD24064 to S.T.W.), the Emory Genetics Discovery Fund, and the Autism Speaks grant (#7660 to X.L.).

AUTHOR CONTRIBUTIONS

T.W., S.T.W. and P.J. designed the study and interpreted the results. T.W. and H.W. analyzed the data. T.W. performed the majority of experiments; Y.L., L.L., X.L. performed 5hmC capture and parts of library preparation. M.Y. C.X.S, H.G. and C.H. assisted with the TAB-Seq experiment and 5hmC capture experiment. A.D. and K.E.S. contributed to the Illumina sequencing, I.G. and K.R. contributed array CGH experiments. I.C., S.C., J.H., M.K., Y.Y., and Q.C. provide some of the hESC and hiPSC lines. T.W., S.T.W. and P.J. wrote the paper with assistance from H.W.

COMPETING FINANCIAL INTEREST

The authors declare no competing financial interests.

Figure 2-1

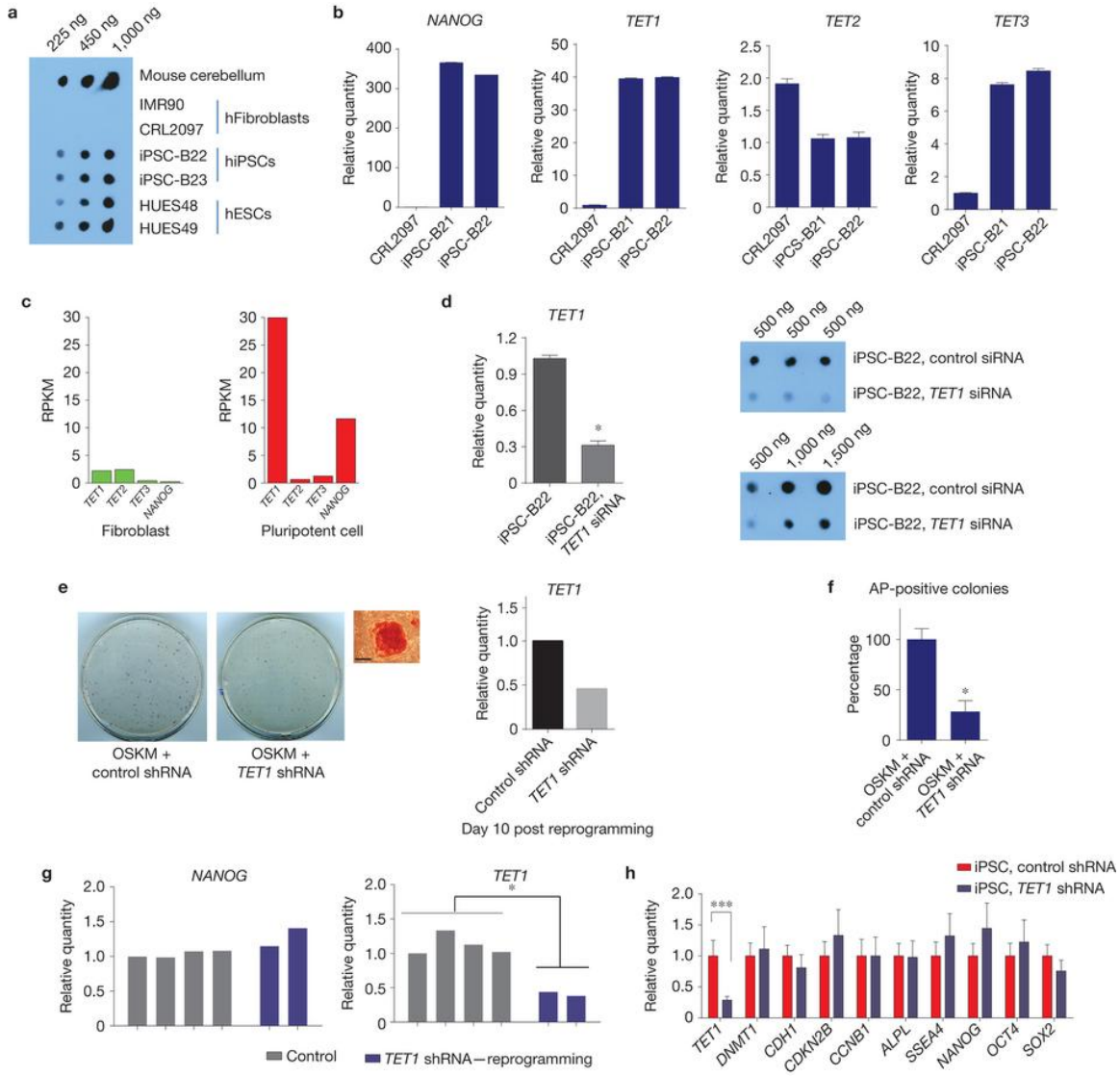

Figure 2-1. TET1 is associated with increased hydroxymethylation during human iPSC reprogramming. (a) Measurement of 5hmC levels in genomic DNAs from fibroblasts, hiPSCs and hESCs by dot blot using anti-5hmC antibody. Mouse cerebellum genomic DNA was used as a control. 225 ng, 450 ng and 1000 ng DNA were used for each sample. **(b)** Quantitative RT-PCR to detect mRNA levels of TET1, TET2, TET3 and NANOG in fibroblasts (CRL2097) and hiPSCs (iPSC-B21, iPSC-B22). Error bars

represent the standard error of the mean (S.E.M.) collected from three independent experiments. **(c)** Boxplot of transcript copy numbers of TET1, TET2, TET3, and NANOG in IMR90 (fibroblasts) and H1 (hESCs) represented by RPKM in RNA-seq. **(d)** Knocking down TET1 by siRNA significantly decreases 5hmC levels in hiPSCs. Left panel represents siTET1 knock down efficiency by quantitative RT-PCR (* t-test, $p<0.05$). Right panel depicts the effect of total 5hmC levels 48hours post siTET1 transfection. Error bars represent S.E.M. collected from three independent experiments. **(e)** Alkaline phosphatase (AP) staining of reprogrammed cells treated either with shTET1 lentivirus or an equal titer shControl lentivirus after O,S,K,M retroviral transduction of 100,000 CRL2097 cells on day 20. Cells used for staining were grown in 10 cm dishes. The image on the right shows a representative AP positive colony and TET1 transcript level in shTET1- or shControl-treated cells 10 days post transduction in one representative experiment of three independent experiments. Scale bars: 300 μm. **(f)** Summary of quantitative analysis of AP-positive colonies in three different experiments (* t-test, $p<0.05$). Controls were normalized to 100%. Error bars represent the standard deviation (SD). **(g)** Real time PCR analysis of TET1 and pluripotency marker NANOG. shTET1-treated reprogrammed colonies maintained normal levels of NANOG, but shows decreased TET1 expression (* t-test, $p<0.05$). Colonies were picked and maintained in puromycin medium (0.5 μg/ml) on puromycin resistant MEFs. **(h)** Real time PCR analysis of normalized gene expression levels of TET1 and selected pluripotency related factors in stable shTET1 or shControl iPS-B22 cells under the puromycin selection (0.5 μg/ml) (*** t-test, $p<0.05$). Error bars represent the S.E.M. of three independent experiments.

Figure 2-2

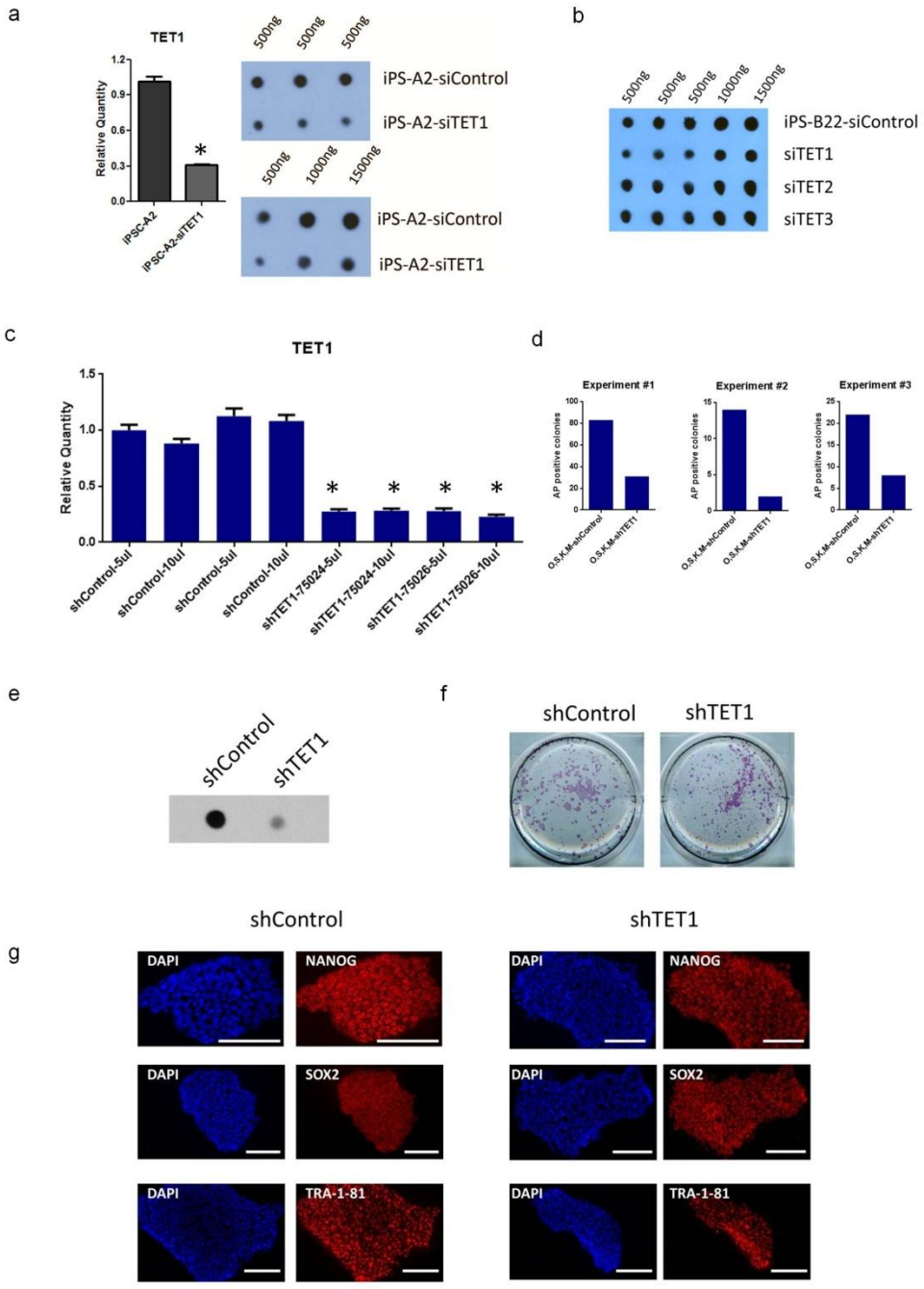

Figure 2-2. TET1 is associated with increased hydroxymethylation during human iPSC reprogramming, but reduction of TET1 does not compromise the

pluripotency of human iPS cells. (a) Knocking down TET1 by siRNA significantly decreases 5hmC levels in iPS-A2. The left panel represents siTET1 knock down efficiency by quantitative RT-PCR (*t-test, $p<0.05$). Error bars represent S.E.M. collected from three independent experiments. Right panel depicts the effect of total 5hmC levels 48hours post siTET1 transfection. **(b)** siTET1 only, but not TET2 or TET3 could affect 5hmC levels in iPSCs. **(c)** shTET1 lentivirus (two shTET1 vectors, 75024 and 75026) could efficiently knock down TET1 (*t-test, $p<0.05$). Error bars represent S.E.M. collected from three independent experiments. **(d)** Quantitative analysis of AP-positive colonies in three different experiments: left, day 20; middle, day25; right, day 20. shGFP lentivirus was used as control for shTET1 lentiviral transduction. **(e)** Dot blot analysis of 5hmC levels in stable shTET1 or shControl iPS-B22 cells. **(f)** Wells stained for alkaline phosphatase for the shControl and shTET1 cells. Cells used for staining were grown in 6-well plate. **(g)** Immunostaining for pluripotency markers NANOG, SOX2 and TRA1-81 in both cell groups (Scale bars: 120 μm).

Figure 2-3

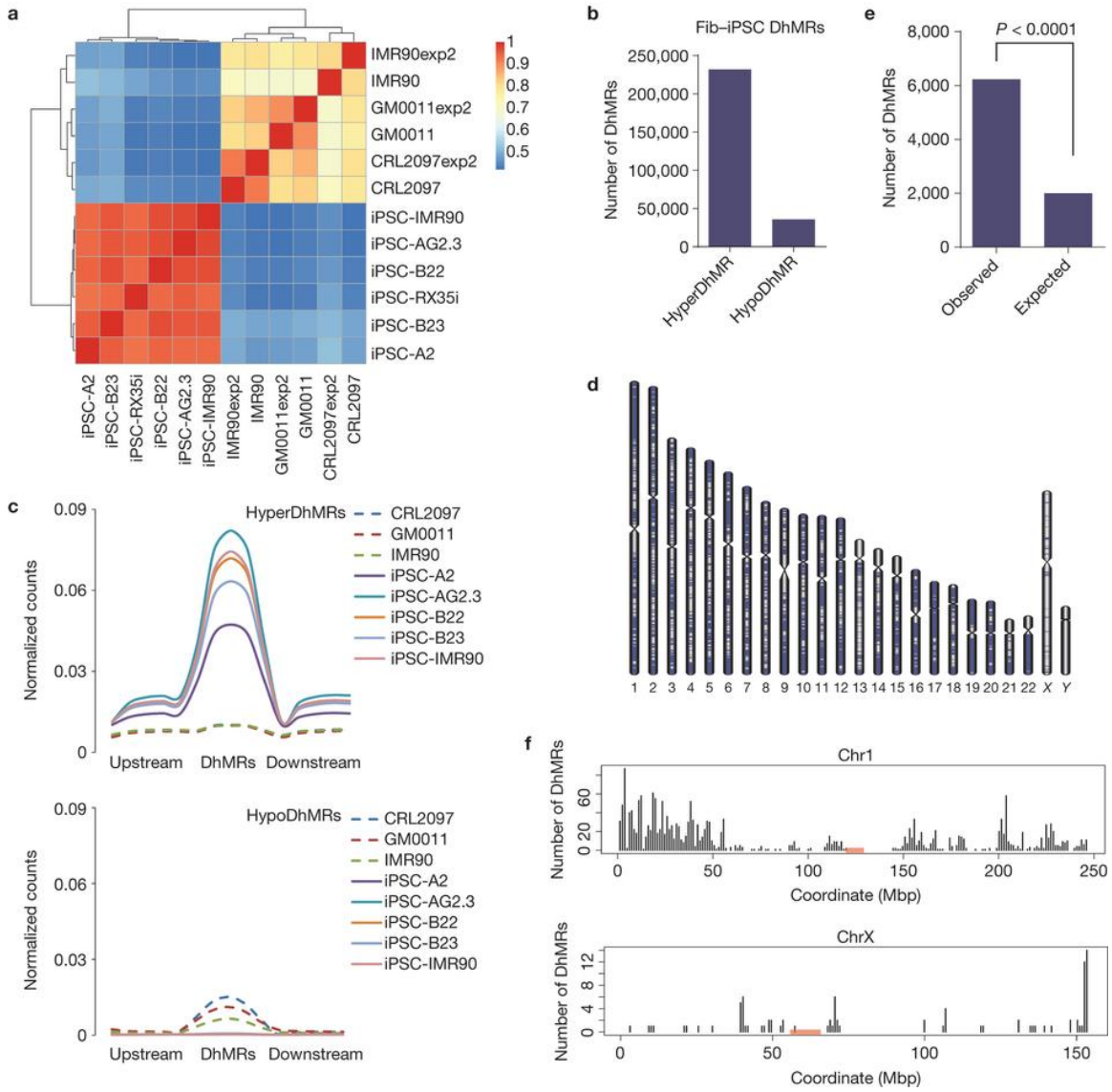

Figure 2-3. Reprogramming confers a 5hmC epigenome in a pattern with a bias towards telomere proximal regions in autosomes. (a) Pearson correlation analysis and cluster among fibroblasts and fibroblast derived iPSCs. The values close to 1 indicate greater similarity. **(b)** Summary of the numbers of 5hmC differentially modified between fibroblasts and iPSCs, indicated by hyperDhMR (iPSCs>Fibroblast) and hypoDhMR (iPSCs<Fibroblast). The regions enriched either in fibroblasts or in iPSCs were subjected

to DhMR calling. 5hmC enriched regions in 3 fibroblast lines and 5 fibroblast-derived iPSC lines were coalesced into a union window. Then the reads in these windows were recounted and normalized to the total read count from the respective cell line. 267,664 DhMRs were called with a FDR of 0.01 by the Bioconductor Deseq package, which uses a negative binomial model for testing differential expression of sequencing data. Among them, 231,866 are hyperDhMRs, and 35,798 are hypoDhMRs. **(c)** Composite 5hmC enrichment profile for fibroblasts and iPSCs in the upstream regions of DhMRs, DhMRs, and downstream regions of DhMRs. The length for upstream and downstream of DhMRs is 5kb. **(d)** Chromosome ideograms showing the genome-wide distribution of the top 20,000 Fib-iPSC-DhMRs ranked by lowest adjusted p-value. Blue lines indicate location of DhMRs. **(e)** Observed and expected numbers of hyperDhMRs occurring at telomere-proximal regions (chi-square test, p value<0.00001). Telomere-proximal regions were defined as regions at either end of a chromosome with a length equal to $1/20^{th}$ of the total length of that chromosome. The observed number occurring at telomere-proximal regions is called by overlapping with top 20,000 hyperDhMRs. The expected number is calculated based on the proportion of total telomere-proximal region length compared to the whole length of all chromosomes. The top 20,000 hyperDhMRs were based on 3 fibroblast lines and 5 fibroblast-derived iPSC lines 5hmC profiles. **(f)** The distribution of the top 20,000 Fib-iPSC-hyperDhMRs in Chr1 and ChrX.

Figure 2-4

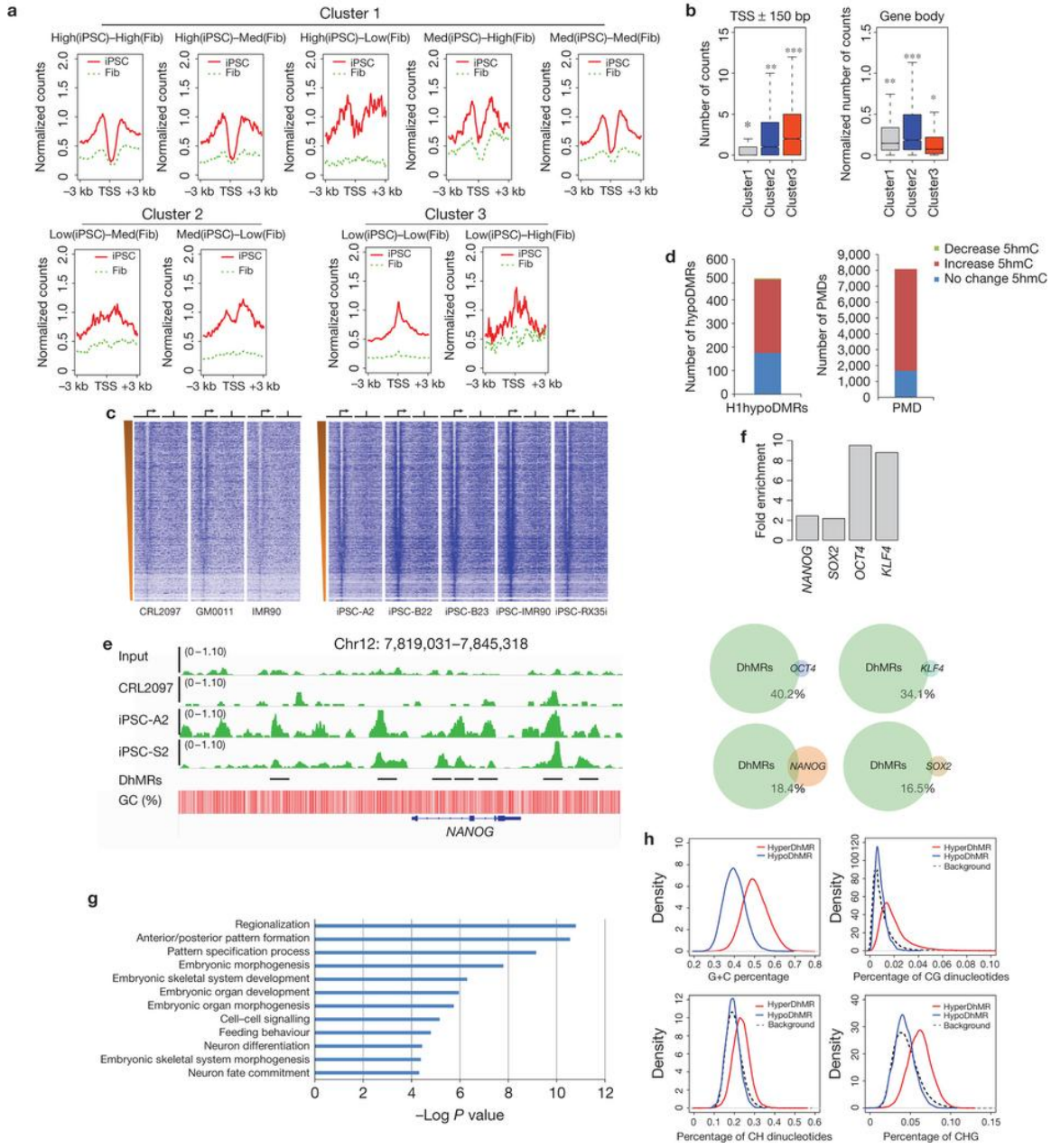

Figure 2-4. 5hmC is associated with gene activity and pluripotency regulatory networks in stem cells. (a) 3 distinct clusters of 5hmC-density pattern at TSS regions (+/- 3kb) in iPSCs and fibroblasts among 9 categories. The 9 categories were classified based on the gene expression changes between iPS cells and fibroblasts: Category 1: high

expression in iPS cells, low expression in fibroblast; Category 2: medium expression in iPS cells, low expression in fibroblast, *etc*. **(b)** Box plots of hydroxymethylation levels in TSS regions and Gene bodies among the three clusters. *** indicates significantly more 5hmC levels compared with all others ($P < 0.001$, Wilcoxon rank test). Similarly, * indicates lowest 5hmC levels, ** indicated intermediate 5hmC level. **(c)** 5hmC enrichment density heatmap. Genes were ordered by expression level from high to low as determined by H1 RPKM[27]. The TSS and direction of transcription of genes are indicated by the genomic region from –3kb to +3kb and an arrow. The TES is indicated by the genomic region from –3kb to +3kb and vertical lines. The left part of the panel shows genes in fibroblasts, the right part shows the genes in iPSCs. **(d)** The correlation between PMD (methylation level is higher in stem cells) and DhMRs, and the correlation between hypoDMRs (methylation level is lower in stem cells) and DhMRs. **(e)** 5hmC density at the NANOG locus in input, iPSCs, and fibroblast cell lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Black lines indicate the DhMRs. **(f)** The overlap between NANOG, OCT4, KLF4, SOX2 binding sites in ES cells and 5hmC significant change regions, shown are observed-to-expected ratios. Lower panel shows the overlapping percentage of each binding sites. **(g)** Gene ontology analysis for genes overlapped with most significant DhMRs. **(h)** Plot of hyperDhMR and hypoDhMR densities in the context of C+G percent, CG percent, CH percent and CHG percent.

Figure 2-5

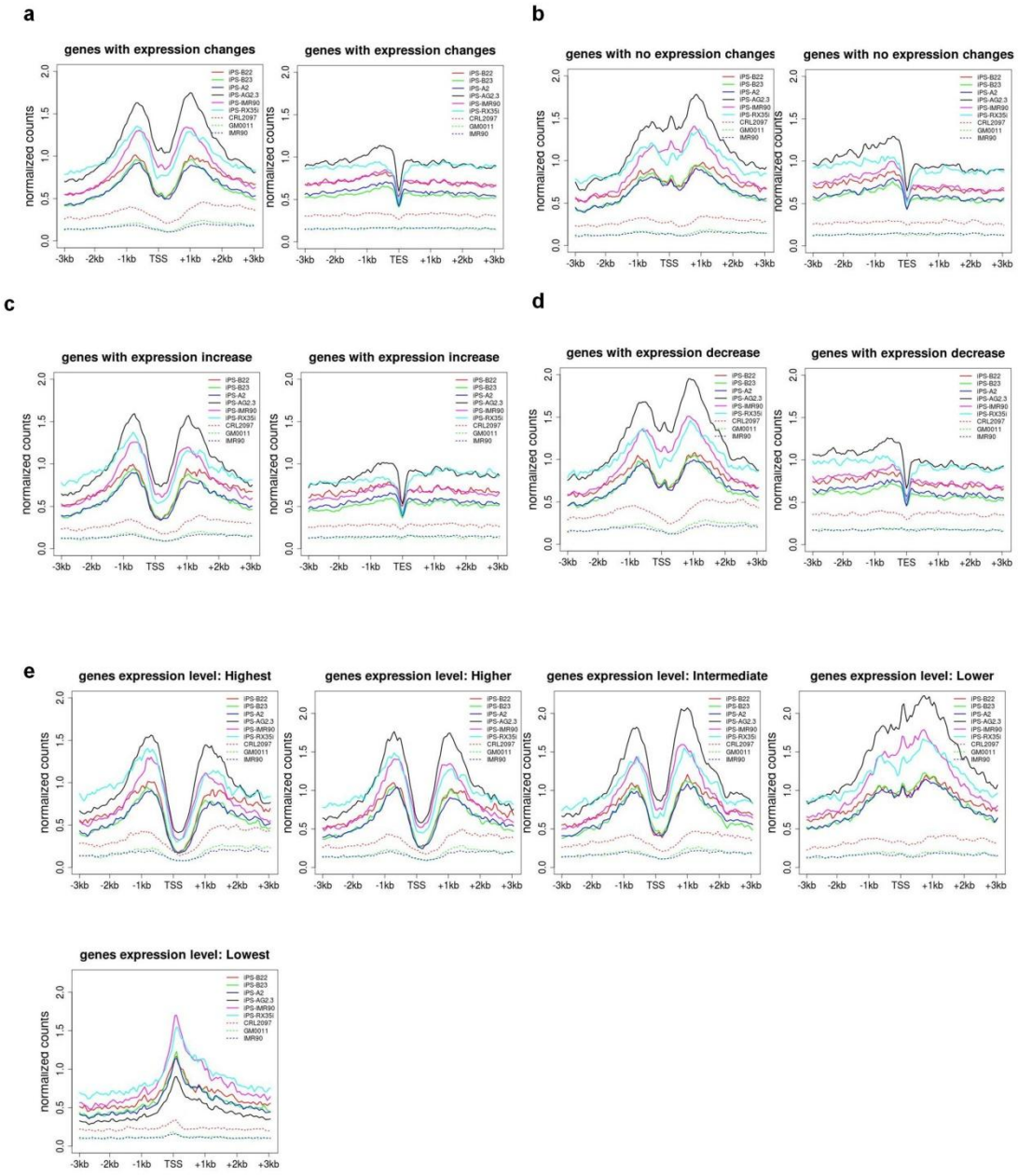

Figure 2-5. Bimodal distribution of 5hmC around TSS and TES. Normalized 5hmC and input read densities among TSS- and TES- surrounding regions. Reads were summed in 50-bp windows 3 kb upstream and downstream of TSS and TES. **(a)** All genes with

expression level changes between iPSCs and fibroblasts interrogated by Affymetrix hg133uplus2 array. **(b)** Genes with no expression changes. **(c)** Genes with expression increased in iPSCs. **(d)** Genes with expression decreased in iPSCs. **(e)** Genes stratified by 5 groups according to H1 RPKM values.

Figure 2-6

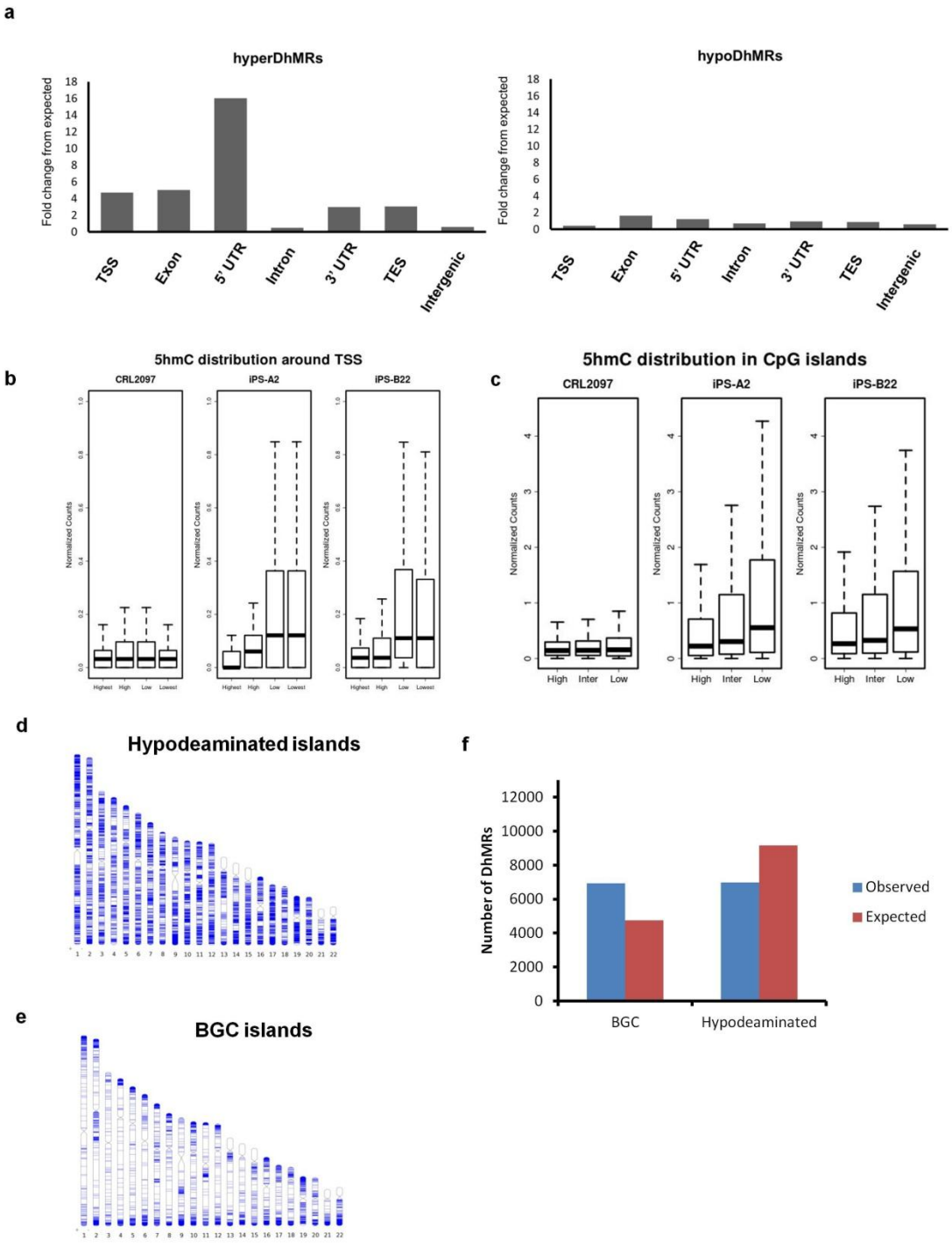

Figure 2-6. Sequence preference of hydroxymethylation modification during reprogramming. (a) Functional annotation of hyperDhMRs and hypoDhMRs between

iPSCs and fibroblasts. TSS represents ±1kb of transcription start site; TES represents ±1kb of transcription end site**. (b)** 5hmC density around core TSS is negatively correlated with gene activity. Genes were divided into four groups (Highest, High, Low, Lowest) according to the RPKM of H1 hESC data. The core TSS region is defined as ±200 bp of transcription start site. **(c)** 5hmC density is negatively correlated with CG dinucleotide percentage in UCSC CpG islands**.** CpG islands are artificially divided into three groups (High, Intermediate, and Low) according to their CG dinucleotide percentages, and then plotted with the 5hmC-normalized counts within these regions. Box shows the center quartile, with the outliers suppressed. **(d)** Chromosomal distribution. A different classification of CpG islands was reported based on the conservation of CpGs across species during evolution. One classification is hypodeaminated CpG islands, which are CpG-rich regions characterized by evolutionarily slow rates of CpG loss, and represent genomic regions with low levels of methylation Shown is the chromosomal layout of CpG-rich loci that were classified as hypodeaminated islands, having low levels of DNA methylation and low deamination rates. **(e)** Chromosomal distribution of Biased Gene Conversion (BGC) islands. Shown is the chromosomal layout of CpG-rich loci that were classified as methylated and hyperdeaminated islands. They exhibit more rapid rates of CpG loss evolutionarily as well as higher methylation levels. The distribution is shown to be highly non-uniform, with hotspots on most subtelomeric regions. **(f)** Fib-ES-DhMRs favors evolutionarily less conserved BGC CpG island groups. The relative Fib-ES-DhMR overlapping regions with BGC and hypodeaminated CpG islands; expected number is also plotted. Of the top 20,000 (lowest adjusted P-values) hyperDhMRs, more regions than expected overlaps with 26,058 BGC islands; in contrast, less regions than expected

are located in hypodeaminated CpGs. This distribution suggests a significant (p-value < 2.2e-16) bias for hydroxymethylation to occur at BGC islands. Therefore, 5hmC modifications acquired during reprogramming tend to occur within the unique sequence context of BGC islands, in which the methylation is evolutionarily less conserved.

Figure 2-7

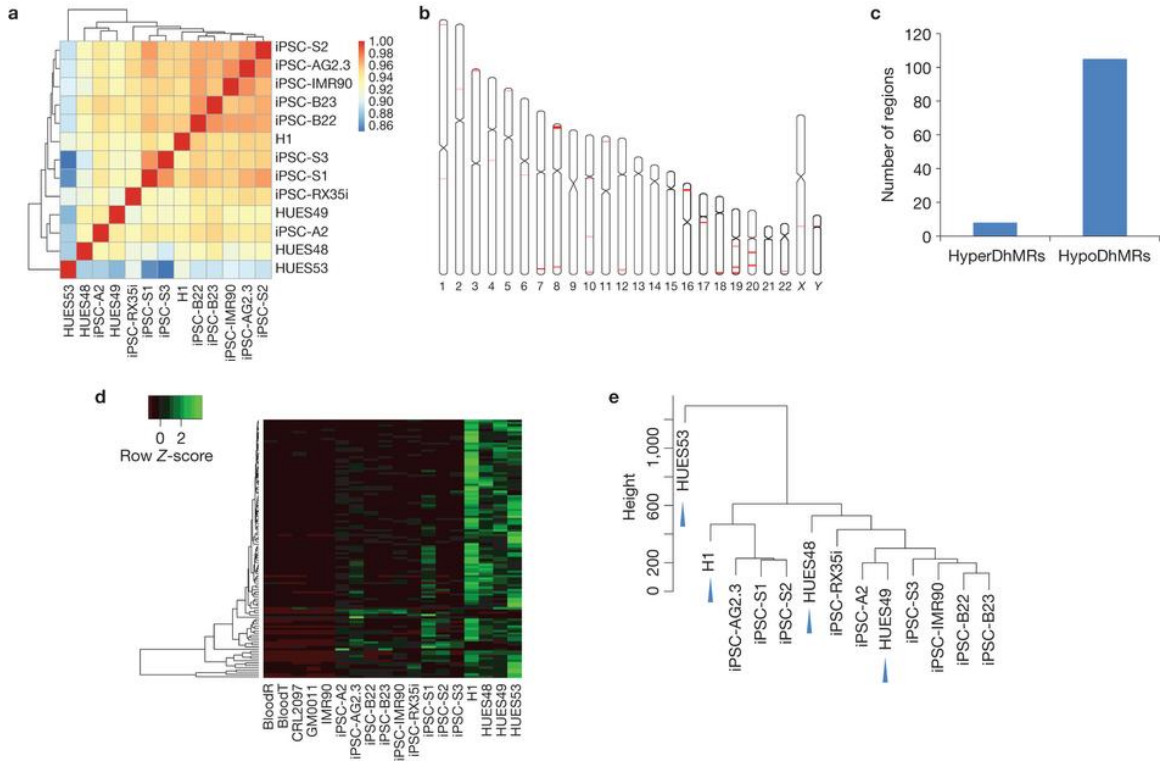

Figure 2-7. Aberrant 5hmC reprogramming hotspots cluster at subtelomeric regions.(a) Pearson correlation analysis and clustering among 9 iPSCs and hESCs. Values close to 1 indicate greater similarity. **(b)** Chromosome ideograms showing the genome-wide distribution of 113 iPSC-ES DhMRs. Red lines indicate locations of DhMRs. **(c)** The number of iPS-ES-hyperDhMRs and iPS-ES-hypoDhMRs. The 372,423 5hmC-enriched regions either in 9 iPSC lines or 4 hESC lines were subjected to DhMR calling by Bioconductor Deseq package. This analysis led to the identification of 113 iPS-ES-DhMRs that were differentially hydroxymethylated in at least one iPS cell or ES cell line (FDR<0.01). 105 of the 113 iPS-ES DhMRs are hypo-hydroxymethylated, with 5hmC levels similar to their respective progenitors. **(d)** Complete linkage hierarchical clustering of 5hmC density within the iPS-ES-DhMRs. The raw count values are scaled

by rows during clustering. **(e)** Hierarchical cluster analysis using the top 1,000 most variable 5hmC enriched regions across all iPSC and hESC samples. Arrow indicates hESCs.

Figure 2-8

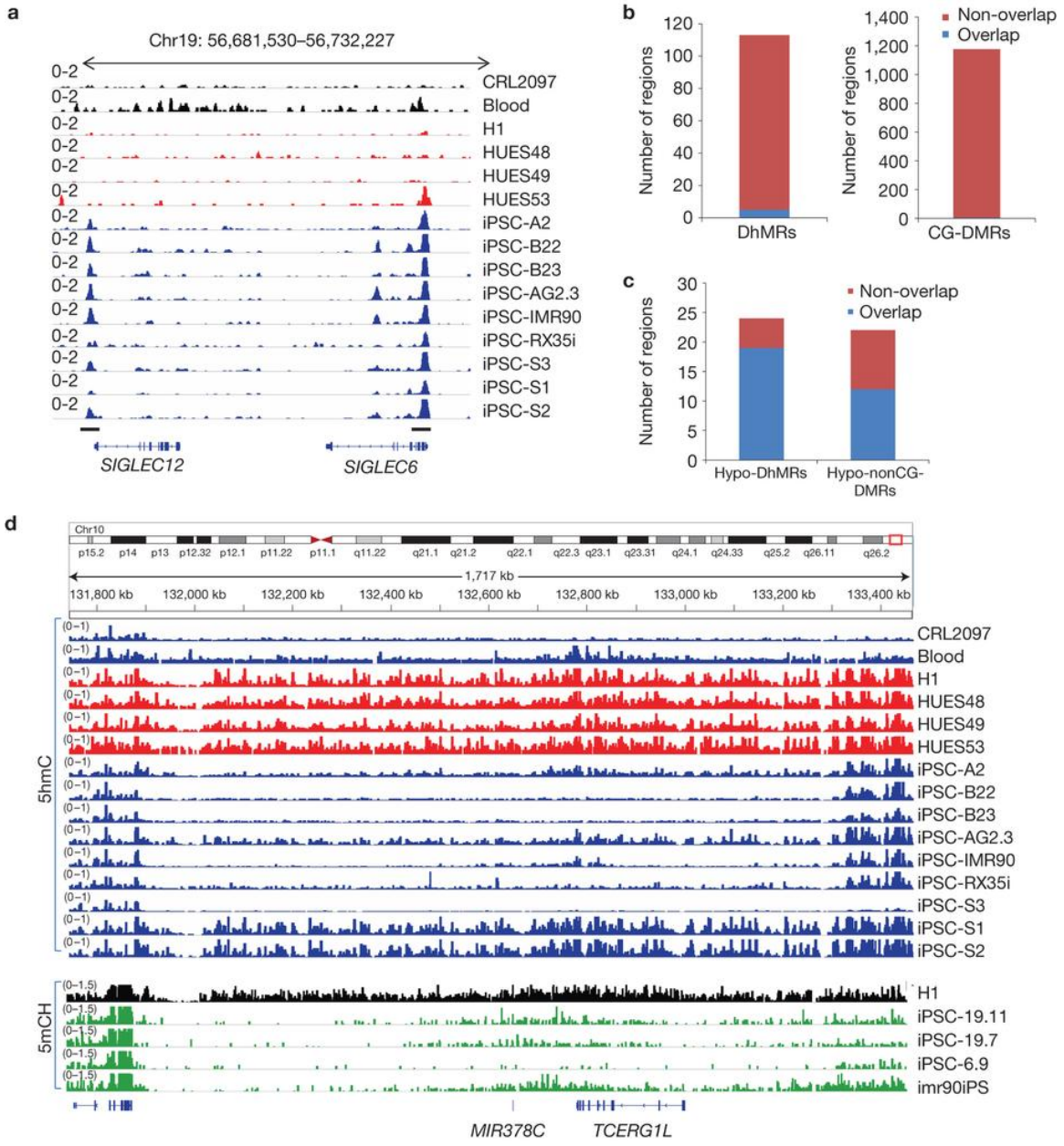

Figure 2-8. 5hmC DhMRs largely overlap with non-CG-DMRs in a large-scale pattern. (a) 5hmC density at the iPS-ES-DhMR SIGLEC6, SIGLEC12 locus, in fibroblast (CRL2097), blood, iPS, and ES cell lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Black bars indicate DhMRs.

(b) The number of 5hmC DhMRs that overlaps with CG-DMRs. CG-DMRs were categorized by methylation state relative to the ES cells. **(c)** The number of 5hmC large-scale hypoDhMRs that overlap with nonCG-DMRs. NonCG-DMRs were categorized by methylation state relative to the ES cells reported previously[17]. The overlap is called if overlapping length is larger than 1 kb. First bar summarizes the overlap for large-scale hypoDhMRs with hypo-nonCG-DMRs. The second bar summarizes the overlap for hypo-nonCG-DMRs with large-scale hypoDhMRs. The blue colour represents overlap between nonCG-DMR and hypoDhMRs. The red colour represents no overlap. **(d)** 5hmC density at of iPS-ES-DhMR TCERG1L locus in fibroblast (CRL2097), blood, iPS, and ES cell lines. The position of the loci within the chromosome and the scale are shown above the gene tracks. Lower parts shows the 5mC levels in CH studied by Lister *et al*, black colour indicates H1 stem cells, green depicts iPSCs.

Figure 2-9

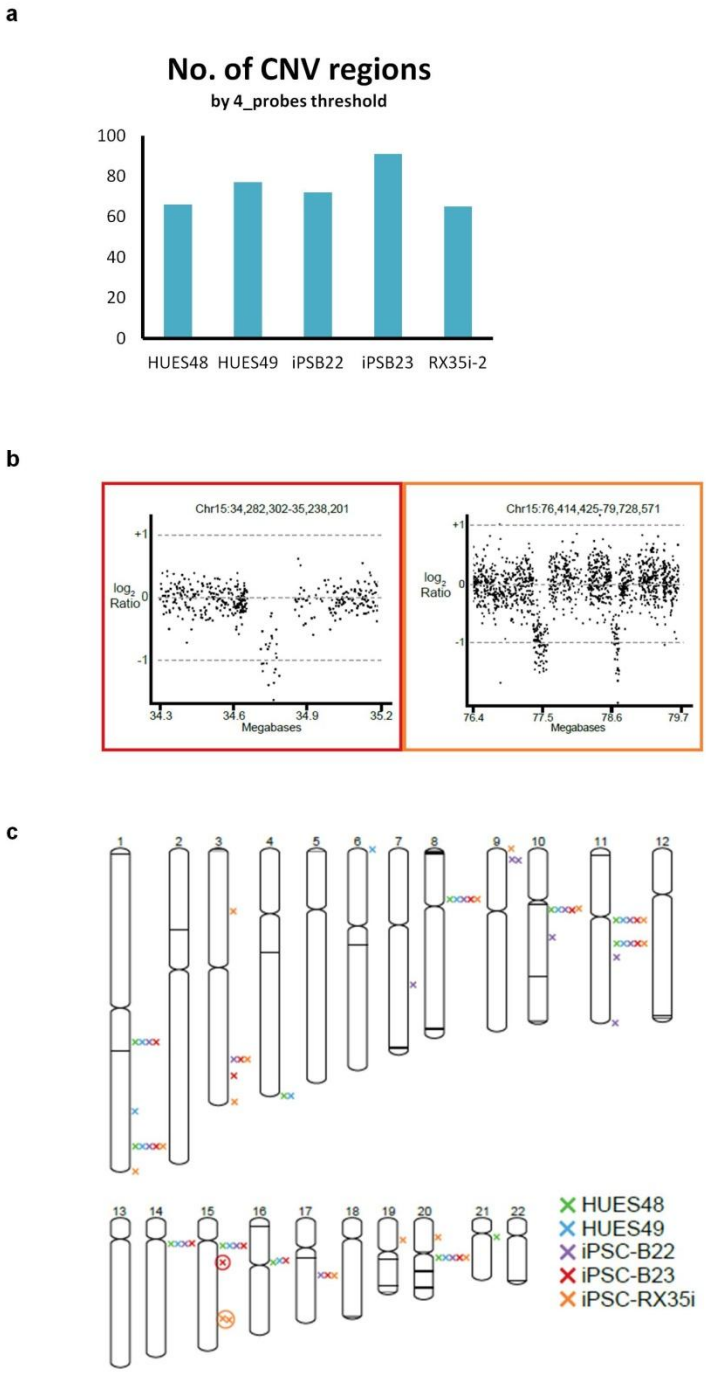

Figure 2-9. The 5hmC aberrant reprogramming hotspots are not due to genomic instability. (a) Number of CNVs called by four consecutive probes with the average of

each probe spanning 3 kb across the genome. **(b)** Density plot of log ratios of signal of two identified CNV regions in iPS-B23 and iPS-RX35i, both of which show a deletion of sequence compared with H1. **(c)** Chromosome ideograms show no overlap of CNVs with 5hmC aberrant reprogramming hotspots. CNV regions shown are indicated by figure annotated bottom right; the hotspot regions are labeled with each individual black line on chromosomes.

Figure 2-10

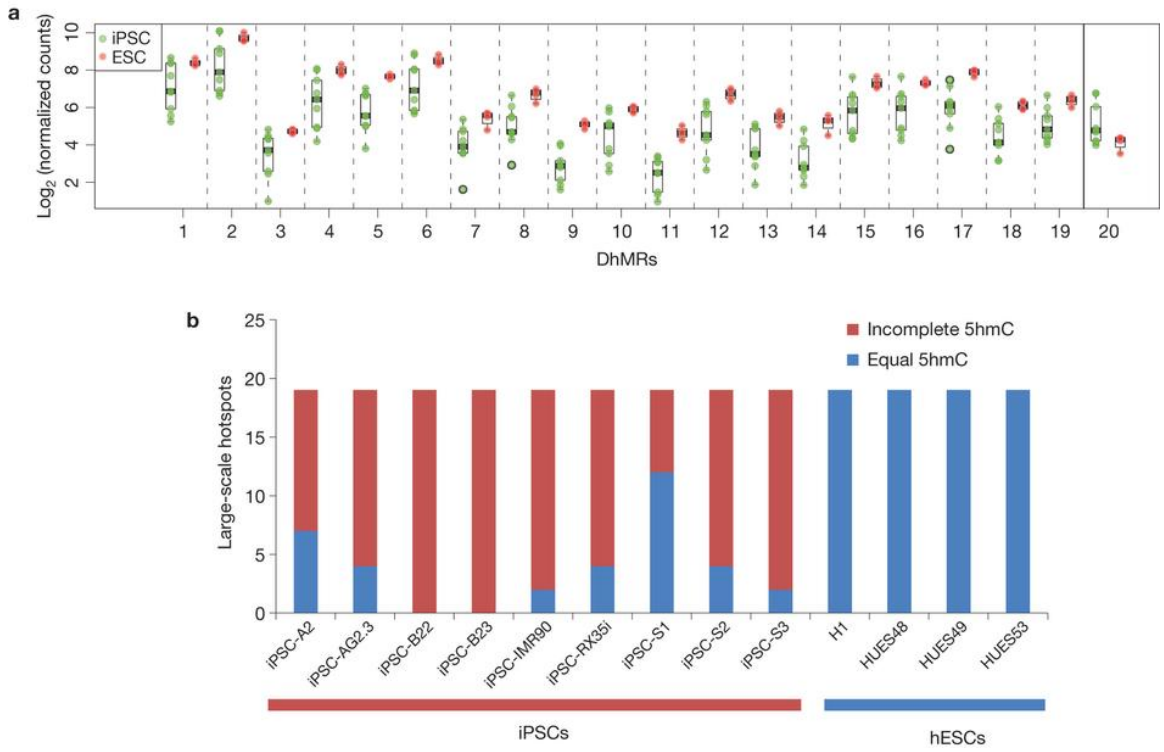

Figure 2-10. Large-scale incomplete hydroxymethylation hotspots are characteristics of human iPS cells. (a) Distribution of 20 5hmC large-scale DhMRs in iPSCs and ESs respectively. Green colour: 9 iPS cell relative enrichment counts, Red colour: 4 hESC cell relative enrichment counts. Solid vertical line separates hyperDhMRs and hypoDhMRs. **(b)** Summary of 19 hypo large-scale DhMRs in each iPSC line. Blue colour indicates regions have similar 5hmC level compared with ES cells, red colour indicates a lower 5hmC level than ES cells. 5hmC levels were determined by counting 5hmC Capture-Seq reads within each hypo large-scale DhMRs for each cell line. A lower 5hmC level in iPS cells is determined by the criteria that 5hmC levels are less than three standard deviations from the mean among ES cells; if levels are within three standard deviations, the region is considered having similar 5hmC levels.

Figure 2-11

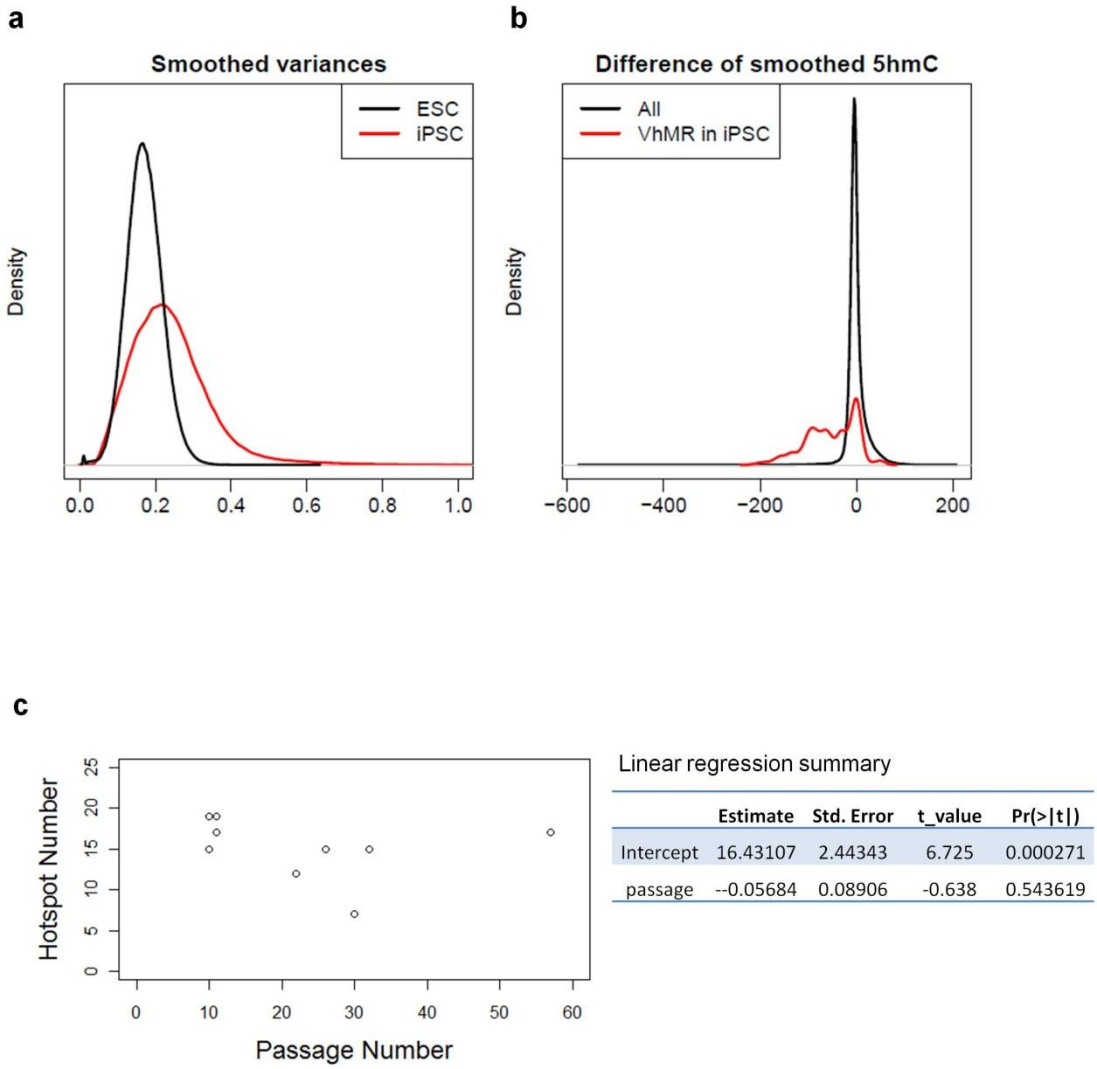

Figure 2-11. Large-scale DhMRs in iPS cells are more variable than in hES cells. (a) Density plot of smooth variance of identified VhMR in ESCs and iPSCs. **(b)** Difference of smoothed 5hmCs. **(c)** Correlation between the number of large-scale DhMRs and passage number in 9 iPSC lines. Table summarizes the linear regression on the parameter

passage number, producing coefficient of passage number with a p-value 0.544, which is not significant.

Figure 2-12

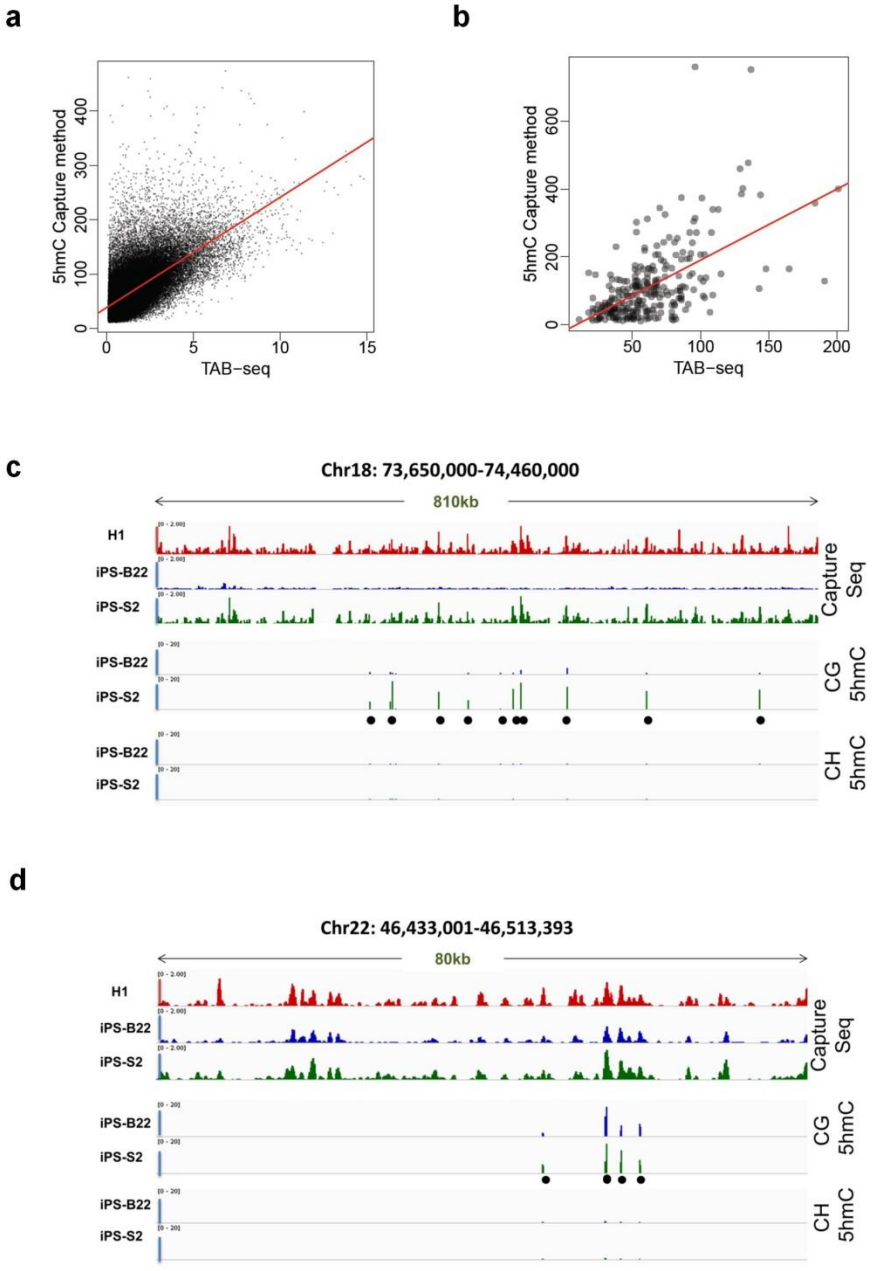

Figure 2-12. Correlation and confirmation analyses between TAB-Seq and 5hmC capture approach. Correlation analysis between TAB-Seq and 5hmC capture

approach in H1 ES cells (a, b). (a) Genome-wide Pearson correlation. Correlation coefficient: 0.65. **(b)** Pearson correlation within 20 large-scale regions. Correlation coefficient: 0.6. The window size used for analysis is 3000 bp. **Chr18 and 22 large-scale hotspots validated at single base resolution by TAB-Seq (c, d). (c)** Summary of PCR-based TAB-Seq in Chr18 large-scale hotspot. The first three tracks are 5hmC intensity determined by capture-Seq, showing iPS-B22 bearing incomplete hydroxymethylation. Below is 5hmC intensity either in CG or CH format determined by PCR based TAB-Seq. Black circles indicate the PCR amplicon mapped loci. **(d)** Summary of PCR based TAB-Seq in Chr22 large-scale hotspot. The first three tracks are 5hmC Capture-Seq results, in some of the regions; both iPS-B22 and iPS-S2 show incomplete hydroxymethylation. In amplified regions by TAB-Seq, both iPS-B22 and iPS-S2 became completely hydroxymethylated. This is also confirmed by TAB-PCR-Seq. Black circles indicate the PCR amplicon mapped loci.

Figure 2-13

Figure 2-13. Large-scale hotspots are caused predominantly by aberrant CpG hydroxymethylation. (a) Summary of PCR based TAB-Seq. **(b)** 5hmC+5mC single base density in one of the amplicons by traditional bisulfite sequencing in 2 hESC and 2 iPSC

lines. Bisulfite sequencing shows the CH methylation (or methylation plus hydroxymethylation) variation in iPS cells. The position of the loci within the chromosome and the scale are shown above the gene tracks. **(c)** 5hmC single base density on CG sites in 15 amplicons by TAB-Seq in 2 human ES cells 4 iPS cell lines. iPS-B22 and B23 shows incomplete CG hydroxymethylation. Green colour indicates iPSCs bearing same hydroxymethylation detected by 5hmC Capture-Seq. Blue colour indicates iPSCs bearing incomplete hydroxymethylation detected by 5hmC Capture-Seq in this region. **(d)** 5hmC+5mC single base density in 15 amplicons by traditional bisulfite sequencing in 2 hESC and 2 iPSC lines. **(e)** 5hmC single base density on CG dinucleotides and CH dinucleotides in one of the amplicons that are marked by blackdot in (c) by TAB-Seq in 2 human ES and 4 iPS cell lines. Green colour indicates iPSCs bearing the same hydroxymethylation detected by 5hmC Capture-Seq. Blue colour indicates iPSCs bearing incomplete hydroxymethylation detected by 5hmC Capture-Seq in this region. **(f)** Schematic summary of large scale incomplete hydroxymethylation on CG dinucleotides in iPS cells.

Figure 2-14

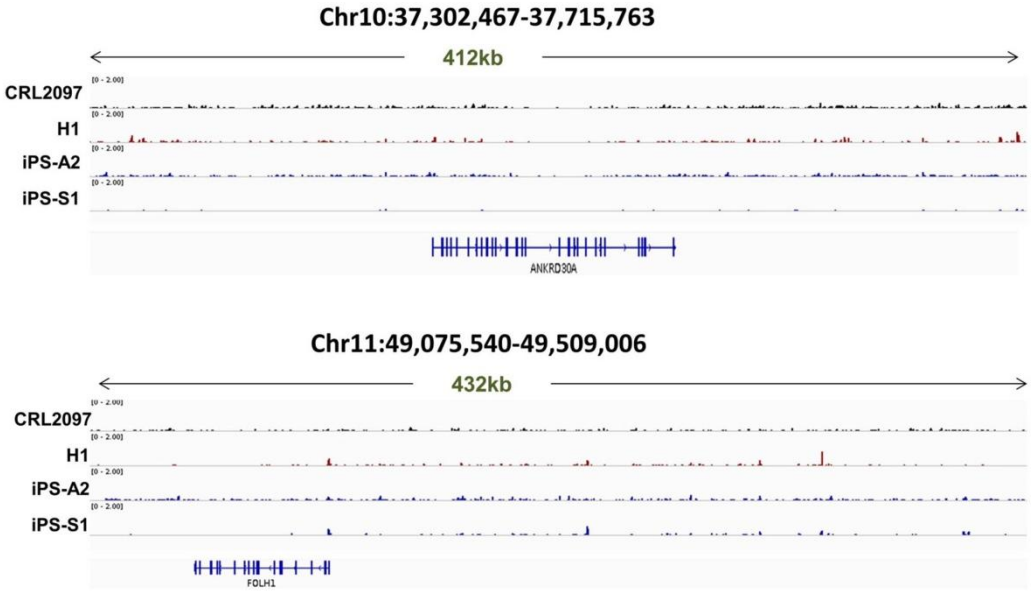

Figure 2-14. Low level of 5hmC at peri-centromeric non-CG DMRs. 5hmC density in fibroblasts, iPS, and ES cell lines at two of the non-CG large-scale DMRs. The position of the loci within the chromosome and the scale are shown above the gene tracks.

Table 2-1. Summary of large-scale hotspots between iPSCs and hESCs

hypoDhMR(19 regions)						
Chr	Range(bp)	Length (bp)	NonCG-DMR	Aberrant Lines No.	Somatic Memory	Genes
Chr1	4533001-5059001	526,001	Y	5	Y	AJAP1
Chr3	474001-592001	118,001	N	9	Y	Intergenic
Chr3	2515001-2907001	392,001	N	7	Y	CNTN4
Chr7	152805001-153016001	211,001	Y	8	Y	Intergenic
Chr7	153184001-153312001	128,001	Y	8	Y	DPP6
Chr7	153461001-153856001	395,001	Y	6	Y	DPP6
Chr7	154010001-154317001	307,001	Y	6	Y	DPP6
Chr8	2681001-3289001	608,001	Y	7	Y	CSMD1
Chr8	138881001-139209001	328,001	Y	7	Y	CSMD1
Chr8	139536001-139818001	282,001	Y	5	Y	FAM135B,COL22A1
Chr10	132010001-133270001	1,260,001	Y	7	Y	TCERG1L,MIR378c
Chr12	125969001-126071001	102,001	Y	5	Y	Intergenic
Chr12	127355001-127814001	459,001	Y	5	Y	TMEM132C
Chr16	6803001-7330001	527,001	Y	5	Y	RBFOX1
Chr18	73780001-74420001	640,001	N	4	Y	Intergenic
Chr20	40395001-40593001	198,001	Y	7	Y	PTRPT
Chr20	41004001-41305001	301,001	Y	7	Y	PTRPT
Chr20	53591001-53742001	151,001	Y	7	Y	Intergenic
Chr22	46433001-46536001	103,001	Y	4	Y	Intergenic
hyperDhMR(1 region)						
Chr	Range	Length	NonCG-DMR	Aberrant Lines No.	Somatic Memory	Genes
Chr22	46005001-46204000	199000	N	6	Y	LOC339685

Table 2-2. Summary of iPSC lines used in this study

	Method	Reprogramming factors	Origin	Teratoma Formation	Passage Number
hiPSC-IMR90	Lenti	O,S,N,L	IMR90	Tested	P57
hiPS-A2	Retro	O,S,K,M	CRL2097	Not tested	P22
hiPS-AG2.3	Lenti	O,S,N,L	Fibroblast	Tested	P26
hiPS-B22	Retro	O,S,K,M	GM0011	Not tested	P10
hiPS-B23	Retro	O,S,K,M	GM0011	Not tested	P11
hiPS-RX35i	Lenti	[OKSM]	Fibroblast	Not tested	P10
hiPS-S1	Sendai	O,S,K,M	Peripheral Blood	Tested	P30
hiPS-S2	Sendai	O,S,K,M	Peripheral Blood	Tested	P32
hiPS-S3	Lenti	[OKSM]	SHED	Not tested	P11

Table 2-3. Summary of 5hmC sequencing statistics

	Total Reads	Unique Alignment	Non Duplicate Reads	Peaks over H1 input
IMR90input	45,312,618	74.89%	27,018,510	-
H1input	29,406,495	69.11%	18,750,895	-
H1	106,833,953	71.95%	11,730,568	197,130
HUES48	46,614,652	53.45%	5,125,254	51,891
HUES49	39,430,026	53.95%	8,541,780	112,026
HUES53	45,901,658	55.51%	3,316,887	47,867
hiPS-IMR90	50,177,110	49.53%	14,467,455	189,423
hiPS-A2	35,090,991	46.48%	11,104,965	101,406
hiPS-AG2.3	53,611,290	54.85%	21,692,047	240,982
hiPS-B22	49,186,918	52.36%	12,051,773	181,285
hiPS-B23	46,420,305	53.33%	12,168,160	176,792
hiPS-RX35i	39,544,303	49.55%	15,295,025	162,837
hiPS-S1	38,636,965	78.13%	27,616,903	217,169
hiPS-S2	41,301,319	77.48%	24,642,117	208,133
hiPS-S3	47,333,331	77.77%	29,415,961	180,499
CRL2097	41,016,247	32.70%	3,640,138	79,723
CRL2097duplicate	40,095,335	77.41%	10,488,812	129,232
GM0011	41,432,564	34.56%	3,088,468	78,679
GM0011duplicate	36,662,588	77.86%	4,927,290	98,403
IMR90	40,238,285	21.52%	1,851,274	20,932
IMR90duplicate	46,006,973	77.45%	6,664,005	94,112

Table 2-4. DhMRs pairwise comparison between fibroblast biological replicates, and between iPSCs and original fibroblasts.

Comparison	Number of 5hmC Peaks
CRL2097_2vs CRL2097	1825
iPSA2vsCRL2097	14233
IMR90_2vsIMR90	1302
iPSIMR90vsIMR90	11282
GM0011_2vsGM0011	3589
iPSB22vsGM0011	29996
iPSB23vsGM0011	18871

Table 2-5. Summary of quantitative RT-PCR primers used in this study.

GAPDH	CATCAATGGAAATCCCATCA	F
GAPDH	GACTCCACGACGTACTCAGC	R
NANOG	AATACCTCAGCCTCCAGCAG	F
NANOG	ACCAGGTCTTCACCTGTTTGT	R
TET1	AATGGAAGCACTGTGGTTTG	F
TET1	ACATGGAGCTGCTCATCTTG	R
TET2	AATGGCAGCACATTGGTATG	F_set1
TET2	AGCTTCCACACTCCCAAACT	R_set1
TET3	ATGTACTTCAACGGCTGCAA	F_set1
TET3	CGGAGCACTTCTTCCTCTTT	R_set1
TET2	GTGAGATCACTCACCCATCG	F_set2
TET2	CAGCATCATCAGCATCACAG	R_set2
TET3	GAGGAGCGGTATGGAGAGAA	F_set2
TET3	AGTAGCTTCTCCTCCAGCGT	R_set2
CDH1	GGTCAAAGAGCCCTTACTGC	F
CDH1	TGGCTCAAGTCAAAGTCCTG	R
CCNB1	GGAAACATGAGAGCCATCCT	F
CCNB1	TTCTGCATGAACCGATCAAT	R
ALPL	CATTGGCACCTGCCTTACTA	F
ALPL	GCTCCAGGGCATATTTCAGT	R
CDKN2B	GGGAGAAGGCAGTGATTAGC	F
CDKN2B	AGCAGACATTGGAGTGAACG	R
OCT4	GAGAAGGATGTGGTCCGAGT	F
OCT4	GTGCATAGTCGCTGCTTGAT	R
DNMT1	CGTTCAACATCAAGCTGTCC	F
DNMT1	CTGCCTTTGATGTAGTCGGA	R

Table 2-6. Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 10 and corresponding amplicon coordinates.

primer	primer_sequence	hg18_coordinates
10_3_F	ATTTGGGGAAGGTTAGGAAATA	chr10:132129447-132129866
10_3_R	CTTCCCCATACTAAAAAAATCAA	
10_5_F	TGTGTTTTAATAATAGAGTGGTTGG	chr10:132167043-132167462
10_5_R	AAAAAAAACAAAATTACCCCC	
10_6_F	TTAGGTGGTTTATTGTGGGAG	chr10:132167326-132167745
10_6_R	CAAAAAAAAACCTACCTTCCC	
10-17F	AGAGGTAGAGTTGTGAGTGTATTTAA	chr10:132772655-132773074
10-17R	CCTCCTAAAAATAAAAAACCAACC	
10-19(3)F	AGGGGTTGGTAGATTTGG	chr10:132775917-132776336
10-19(3)R	AAAACCCTTACCTTTCACTAACA	
10-21F	AGAGTTGTTAGGTTAGGTGGTGTTT	chr10:132782192-132782611
10-21R	TAATCCCTCTACCTCTATCCCTATC	
10-26(1)F	TGGGAAGAGTTGAAGTTTTTTT	chr10:132814608-132815027
10-26(1)R	CCTAACAACACATAACTACCCTACC	
10-29(1)F	TATTGGGTTTTTTGGTYGTTT	chr10:132825447-132825866
10-29(1)R	ACTCCCATACCTTCTTCCAAAC	
10-30(1)F	GGAAGGAAAGGAAAAAAGTTTTT	chr10:132835493-132835912
10-30(1)R	TAAAAATTAAACCCAAAACCCC	
10-31(1)F	GGGAGGTTTTGTTAGGAATAGG	chr10:132851454-132851873
10-31(1)R	CTATCCCAAAAACCTCAAACC	
10-32(1)R	CCCCATAAACCCTAATATACCA	chr10:132863319-132863738
10-32(2)F	TTGGAGATGTAGGAAGTGTTGT	
10-33(1)F	ATTGTTTGTAGAGTTTGTGGTTTT	chr10:132919649-132920068
10-33(1)R	TTAACTTCATCTCTCTCCCACAA	
10-36(1)F	TGGGAGAGTTGGGGTTTAG	chr10:132939228-132939647
10-36(1)R	TCAAAACCACRACAAAACTC	
10-37(1)F	GTTGTTTTTGGAGTTAGGGGA	chr10:132945333-132945752
10-37(1)R	CTAATCCCTACTACCCTCCCA	
10-38(1)F	TAAGGAAATTATTAGGGATGAGGTG	chr10:132950132-132950551
10-38(1)R	TCCCATTCAAACAAACAAAATA	

Table 2-7. Primers used for PCR-based TAB-Seq targeting large-scale hotspot in chromosome 18 and 22, and corresponding amplicon coordinates.

primer	primer_sequence	hg18_coordinates
22-1F	ATTTTTGTTATAGGGTGTGTGA	chr22:46480570-46480977
22-1R	AATTAAACTAACACAAACCACCT	
22-2F	GTAAGTTGAATTTATGAGAAGGGG	chr22:46488307-46488614
22-2R	CRCTAAAAAACTAAATACATTCCC	
22-3F(2)	TTTGAGTTAGGTTGGGGTTT	chr22:46490087-46490494
22-3R(2)	TCCTCTTCCCACTACAAAAAAC	
22-4F	TGGTGTGGATATAGTGTGTTGT	chr22:46492463-46492878
22-4R	ATTTTCCAAAAATCCCTAAAACT	
22-5F	TTTGTTGAGGATTGGTGG	chr22:46550744-46551051
22-5R	CAACCTAATACTCACCTACTAAAAAA	
18-4F	GAGTTTAGGATTTTGTGAGTTAGG	chr18:73916409-73916780
18-4R	TAAAAAATCCCTACCTAACCCT	
18-5F	GTTAGGGGGATTAGGTGGAT	chr18:73940950-73941421
18-5R	ATTAAAAAACTTAAAAAAATTCCACC	
18-6F	TAGGAGGTTTAGTGGGAAGGT	chr18:73943768-73944239
18-6R	TCCTTCCAAAATCACAACTAAA	
18-8F	GAATGTTGGGTTTTGGATGTT	chr18:73948431-73948702
18-8R	AAAACTCCCTCAAATCCACAAT	
18-10F	GTTGATGGTGGGGATTTTAT	chr18:73999869-74000204
18-10R	CCATCTACTACAAATAACCCCAA	
18-19F	GTATTTGGTGGAGGAGGG	chr18:74035387-74035758
18-19R	CCAACAACAATCATCCTAAAAT	
18-20F	GGGTTTGAATTTAGGATTTTGT	chr18:74035882-74036117
18-20R	TCCTATTCCAAAAACTCTTCCT	
18-21F	AGGAGAGTTGTGGGTTTATTAAA	chr18:74074470-74074841
18-21R	ACACCTATACACACACACAAACA	
18-23F	GATAGGGAATAAGGTGGTTTTA	chr18:74089857-74090328
18-23R	TCTAATTAACATCCTACACCCC	
18-24F	GTTTGTAGAGYGTGGAGAGTT	chr18:74099672-74100143
18-24R	ACCTCCTTCCTTCTTCTCAAA	
18-33F	GTGTTGGTTGTGTTAAGATGATT	chr18:74155400-74155871
18-33R	AAAAAACCCAAACCCTTTC	
18-34F	GGAGAAGTTAGGATTTAGGAGAAG	chr18:74251387-74251858
18-34R	TACAACACTAAACTCTACATCCRA	
18-35F	GGTGTAGGAAGATGAGTTGGTT	chr18:74251569-74252040

18-35R	AAACAAAATCTTCCTCAAACAAA	
18-38F	GGAAGGATAGTGGTTTAGGTTT	chr18:74388487-74388958
18-38R	AAAAAAATTAAACCCAACCC	

REFERENCES

1. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
3. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).
4. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935 (2009).
5. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929-930 (2009).
6. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-1133 (2010).
7. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-843 (2010).
8. Koh, K.P. *et al.* Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell stem cell* **8**, 200-213 (2011).
9. Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications* **2**, 241 (2011).
10. Dawlaty, M.M. *et al.* Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell stem cell* **9**, 166-175 (2011).

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

12. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one* **5**, e15367 (2010).

13. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397 (2011).

14. Szulwach, K.E. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS genetics* **7**, e1002154 (2011).

15. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes & development* **25**, 2436-2452 (2011).

16. Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439-452 (2011).

17. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68-73 (2011).

18. Robinton, D.A. & Daley, G.Q. The promise of induced pluripotent stem cells in research and therapy. *Nature* **481**, 295-305 (2012).

19. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one* **5**, e8888 (2010).

20. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**, 423-434 (2011).

21. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).

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

23. Szulwach, K.E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature neuroscience* **14**, 1607-1616 (2011).

24. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348 (2011).

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

26. Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & development* **25**, 679-684 (2011).

27. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

28. Xu, Y. *et al.* Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell* **42**, 451-464 (2011).

29. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).

30. Cohen, N.M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773-786 (2011).

31. Kim, K. *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285-290 (2010).

32. Kim, K. *et al.* Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. *Nature biotechnology* **29**, 1117-1119 (2011).

33. Polo, J.M. *et al.* Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nature biotechnology* **28**, 848-855 (2010).

34. Hussein, S.M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58-62 (2011).

35. Laurent, L.C. *et al.* Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture. *Cell stem cell* **8**, 106-118 (2011).

36. Chin, M.H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell stem cell* **5**, 111-123 (2009).

37. Guenther, M.G. *et al.* Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell stem cell* **7**, 249-257 (2010).

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

39. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell stem cell* **4**, 487-492 (2009).

40. Wang, W. *et al.* Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog

1. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18283-18288 (2011).

41. Hanna, J.H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143**, 508-525 (2010).

42. Dawlaty, M.M. *et al.* Combined deficiency of tet1 and tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Developmental cell* **24**, 310-323 (2013).

43. Costa, Y. *et al.* NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature* **495**, 370-374 (2013).

44. Onder, T.T. *et al.* Chromatin-modifying enzymes as modulators of reprogramming. *Nature* **483**, 598-602 (2012).

45. Liang, G., He, J. & Zhang, Y. Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming. *Nature cell biology* **14**, 457-466 (2012).

46. Daley, G.Q. *et al.* Broader implications of defining standards for the pluripotency of iPSCs. *Cell stem cell* **4**, 200-201; author reply 202 (2009).

47. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175-181 (2010).

48. Dyachenko, O.V., Schevchuk, T.V., Kretzner, L., Buryanov, Y.I. & Smith, S.S. Human non-CG methylation: are human stem cells plant-like? *Epigenetics : official journal of the DNA Methylation Society* **5**, 569-572 (2010).

49. Okita, K. & Yamanaka, S. Induced pluripotent stem cells: opportunities and challenges. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **366**, 2198-2207 (2011).

50. Carey, B.W. *et al.* Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell stem cell* **9**, 588-598 (2011).

51. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175-181 (2010).

52. Somers, A. *et al.* Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. *Stem Cells* **28**, 1728-1740 (2010).

53. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**, R25 (2009).

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

55. 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).

56. Carvalho, B.S. & Irizarry, R.A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **26**, 2363-2367 (2010).

57. Robinson, J.T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).

Chapter 3

Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum

This manuscript has been published on ***Human Molecular Genetics***. Tao Wang, Qian Pan, Li Lin, Keith E Szulwach, Chun-Xiao Song, Chuan He, Hao Wu, Stephen T. Warren, Peng Jin*, Ranhui Duan*, and Xuekun Li*. Human Molecular Genetics, 2012, 10.1093/hmg/dds394

ABSTRACT

5-hydroxymethylcytosine (5-hmC) is a newly-discovered modified form of cytosine that has been suspected to be an important epigenetic modification in neurodevelopment. While DNA methylation dynamics have already been implicated during neurodevelopment, little is known about hydroxymethylation in this process. Here we report DNA hydroxymethylation dynamics during cerebellum development in the human brain. Overall, we find a positive correlation between 5-hmC levels and cerebellum development. Genome-wide profiling reveals that 5-hmC is highly enriched on specific gene regions including exons and especially the untranslated regions (UTRs), but it is depleted on introns and intergenic regions. Furthermore, we have identified fetus-specific and adult-specific differentially hydroxymethylated regions (DhMRs), most of which overlap with genes and CpG island shores. Surprisingly, during development, DhMRs are highly enriched in genes encoding mRNAs that can be regulated by fragile X mental retardation protein (FMRP), some of which are disrupted in autism, as well as in many known autism genes. Our results suggest that 5-hmC-mediated epigenetic regulation may broadly impact the development of the human brain, and its dysregulation could contribute to the molecular pathogenesis of neurodevelopmental disorders.

INTRODUCTION

Methylation at the 5-positon of cytosine (5-mC), which is catalyzed by DNA methyltransferases (DNMTs), plays important roles in mammalian neuronal systems (1). The proper establishment of DNA methylation is critical for embryonic and postnatal development (2, 3). DNMT3a is required for maintaining neural stem cell self-renewal, and loss of the protein significantly impairs postnatal neurogenesis (4), suggesting a regulatory role of DNA methylation in neurodevelopmental process (5). Dnmt1 and Dnmt3a depletion induces abnormal neuronal phenotypes, including learning and memory deficits and abnormal synaptic plasticity (6). Furthermore, pharmacological inhibition of Dnmt activity can block hippocampus-dependent memory formation (7). Mutation of methyl-CpG binding protein 2 (MECP2), which binds to methylated DNA and acts as a transcriptional repressor or activator, causes Rett syndrome, and related neurodevelopmental disorders (8). These observations indicate that the proper establishment and maintenance of DNA methylation is essential for normal development and function of the mammalian brain.

5-hydroxymethylcytosine (5-hmC), which is converted from 5-mC by ten-eleven translocation (TET) proteins, is present in the mammalian genome (9-11). All TET family proteins can catalyze the conversion of 5-mC to 5-hmC. TET1 was first reported to have a role in maintaining the pluripotent state of embryonic stem cells (ESCs) (12), and together with TET2, it also regulates the cell lineage commitment of ESCs (13). TET2 modulates the balance between self-renewal and differentiation in hematopoietic

stem cell, making it critical for normal myelopoiesis; TET2 mutations are seen in multiple types of leukemia (14-18). TET3 contributes to the global DNA methylation erasure during the zygote stage of embryonic development (19, 20). Taken together, these studies point to a critical role for TET-mediated 5-hmC modification in developmental processes and the possibility that dysregulation of 5-hmC may be associated with disease.

Although DNA methylation has generally been regarded as a highly stable epigenetic mark, recent studies have uncovered DNA methylation changes during brain development and aging, suggesting that epigenetic changes like 5-mC could function as an intermediate step for the internal or external environmental regulation of the brain genome. Studies from our own and other groups have identified strong enrichment of 5-hmC in mammalian brains (9, 10, 21). However, compared with 5-mC, little is known about the roles of 5-hmC in the mammalian brain. Our previous study revealed the dynamics of DNA hydroxymethylation during postnatal development and ageing in mouse brain (21, 22). Another study found that 5-hmC is enriched at promoters and gene bodies, and its enrichment on gene bodies is positively correlated with gene expression in the frontal lobe tissue of human brain (23). Nevertheless, the features of 5-hmC during human brain development remain a mystery.

Here we extend our previous work to profile the genome-wide distribution of 5-hmC during cerebellum development in human fetal and adult brains. We find that the overall

5-hmC level increases during cerebellum development. Most differentially hydroxymethylated regions (DhMRs) between fetus and adult overlap with genes, and are strongly associated with CpG island shores. Strikingly, these 5-hmC changes are highly enriched in genes whose transcripts can be regulated by FMRP, as well as in many genes linked with autism.

RESULTS

Dynamics of DNA hydroxymethylation and its genomic features in human cerebellum

Previous studies indicated 5-hmC levels in mouse cerebellum are high and can change during developmental processes (10, 21). To extend this work, here we focused on the human cerebellum to study the features of 5-hmC in the developing human brain. We first performed 5-hmC specific dot blot with fetus and adult cerebellum DNA samples, and found that the total levels of 5-hmC increased significantly (42.1% increase) from fetus to the adult stage (Fig. 3-1A and B), consistent with the observation in mouse cerebellums.

To profile the genome-wide distribution of 5-hmC in human cerebellum, we isolated genomic DNA containing 5-hmC using chemical capture technique developed previously (22) and then sequenced those DNA fragments. Genome-wide correlation analysis showed that the 5-hmC profiles from two adult cerebellums are more similar to each other than either is to fetal cerebellum (Fig. 3-1C). We then determined the genomic features associated with 5-hmC-enriched regions and found that 5-hmC is highly enriched at genes (Fig. 3-2), particularly 5'-UTR and exons, but it is depleted at introns and intergenic regions in both fetal and adult samples (Fig. 3-1D). Furthermore, 5-hmC is strongly associated with CpG islands and CpG island shores ($p<0.001$, comparing observed and expected frequencies) (Fig. 3-1E). Together, these data suggest that 5-hmC

is significantly increased during the development of human cerebellum and is associated with specific genomic regions.

5-hmC genomic features during cerebellum development

To examine the genomic features of 5-hmC during the development of human cerebellum, we used a Poisson-based model calling method to determine the fetus-specific DhMRs (fetus has higher 5-hmC levels than adult) and adult-specific DhMRs (adult has higher 5-hmC levels than fetus). Across the genome, we identified 28015 DhMRs between the fetus and adult samples, of which 15,829 are adult-specific and 12,186 are fetus-specific DhMRs (Fig. 3-3A). Hierarchical clustering of the top 500 most significant DhMRs indicated a greater similarity between adult samples, and significant difference between fetus and adult, most showing an increase of 5-hmC levels in adult (Fig. 3-3B). We noted that the fold-change of adult-specific DhMRs is much greater than fetus-specific DhMRs (Fig. 3-3C). The majority of DhMRs we identified also showed a strong bias towards CpG islands and CpG island shores ($p<0.001$) (Fig. 3-3D and E). Fetus-specific DhMRs had a stronger tendency to be associated with CpG islands (10.4% vs 5.2%) and CpG shores (22% vs 14%) than adult-specific DhMRs. Furthermore, both fetus- and adult-specific DhMRs overlapped more with CpG shores than with CpG islands (Fig. 3-3D and E).

To examine the potential relationship between DhMRs and gene function, we aligned the identified DhMRs to the annotated human genes and found that 69% (8,407/12,186) of

fetus-specific DhMRs overlapped with genes (Fig. 3-4A) and 72% (11,396/15,829) of adult-specific DhMRs overlapped with genes (Fig. 3-4B). Both fetus-specific and adult-specific DhMRs are also largely localized to UTR and exons, but are depleted at introns and intergenic regions (Fig. 3-4C and D). A similar pattern has been seen in DhMRs identified in mouse cerebellum (21). Intriguingly, fetus-specific DhMRs are more strongly associated with transcription start sites (TSS) than adult-specific DhMRs (Fig. 3-4C and D). Furthermore, around half of the genes associated with DhMRs also show differential hydroxymethylation during cerebellum development in mouse (observed to expected ratio: 1.45, 3.46 and 6.83 for genes that are associated with at least 1, 2, and 4 DhMRs) (Figure 3-4E). Gene ontology analysis indicated that adult-specific DhMRs are enriched at genes involved in ion channel binding and cell-cell adhesion, and fetus-specific DhMRs preferentially localize at genes associated with ion channel and neuronal development (Fig. 3-5A and B). Fig. 3-4E shows DhMRs in two genes, MYOD1 and MAP1B, both of which encode proteins involved in neuronal function.

The fetus-specific DhMRs displays pluripotent epigenetic memories

Epigenetic properties of human embryonic stem cells (hESCs) dictate proper fetal development, and presumably, hESCs share more epigenetic features with the fetus than with the adult (24, 25). Previous studies have also found 5-hmC enriched at genes (26, 27). We compared 5-hmC profiling between H1 hESCs and human cerebellum, and found that 5-hmC is dramatically depleted at TSS in human brain relative to ESCs, but still, retain a bimodal distribution (Fig. 3-6A). Moreover, 5-hmC exhibits less enrichment

at CpG islands and shores in human brain than in hESCs (Fig. 3-6B, C and D). These results indicate that the genomic distribution of 5-hmC has a different pattern in human brain, consisting mainly of post-mitotic cells, than in proliferating hESCs.

We next compared DhMRs identified in human brain with H1 ESCs 5-hmC enrichment peaks that we previously identified (28). We found that 57% of fetus-specific DhMRs were shared with hESCs, but only 19% of adult-specific DhMRs were found in hESCs (Fig. 3-6E). Furthermore, 338 of the 500 most highly enriched fetus-specific DhMRs were shared with ESCs while only 146 of the 500 most highly enriched adult DhMRs occurred in ESCs (Fig. 3-6F). These data indicate the fetus shares more epigenetic memory with ESCs than does the adult; this could play a role in development-specific gene expression, as suggested by the gene ontology analysis of fetus- specific DhMRs described above.

DhMRs are associated with genes involved in neurodevelopmental disorders

DNA methylation-mediated epigenetic modulation plays important roles in neurodevelopment. Dysregulation of DNA methylation can cause disorders such as Rett syndrome and fragile X syndrome (FXS) (1, 29, 30). Recent studies have revealed 5-hmC-mediated epigenetic dynamics during embryonic (19, 20) and postnatal development (21), which indicates a potential function for 5-hmC in development and disease (31). More interestingly, both loss- and over-expression of MeCP2 not only affects the overall level of 5-hmC, but also modulates its distribution in the genome

during mouse cerebellum development (21). These studies pointed to a potential role for 5-hmC in neurodevelopmental disorders.

To better understand the roles of 5-hmC in neurodevelopmental disorders, here we asked whether 5-hmC-enriched regions we identified were associated with neurodevelopmental disorders such as FXS, and more broadly, autism spectrum disorder (ASD). We first compared 5-hmC enrichment on all UCSC RefSeq genes and FMRP target genes that were identified previously (32); FMRP targets displayed more 5-hmC enrichment than the RefSeq genes (Fig. 3-7A). We then stratified the RefSeq genes into three groups: highly, moderately and weakly expressed genes based on the RNA expression level in human cerebellum. FMRP target genes displayed more 5-hmC enrichment than any of the three groups ($p<0.001$, Wilcoxon rank test). To rule out the possibility that enrichment of hydroxymethylation in FMRP target genes is a general characteristic of RNA binding proteins in cerebellum, we also looked at TDP-43, which is involved in pre-mRNA splicing and translational regulation of its RNA ligand. Compared with FMRP target genes, we did not see strong 5-hmC enrichment in TDP-43 target genes (Figure 3-7A). We then analyzed 5-hmC mapping reads on all RefSeq genes, highly expressed genes, TDP-43 target genes and FMRP target genes and found 5-hmC showed higher mapping reads on FMRP target genes in both fetus and adult brain versus all other groups ($p<0.001$, Wilcoxon rank test) (Fig. 3-7B, C). Moreover, DhMRs also had a stronger tendency to localize on FMRP target genes than any other group (Fig. 3-7D, E).

FXS is one of the leading monogenic causes of ASD, with up to 30% of FXS patients showing autistic symptoms. Two recent studies found an unusual coincidence between autism-related genes and FMRP target genes, which had more than expected *de novo* mutations in children with ASD (32, 33). Since we saw that 5-hmC significantly overlaps with FMRP target genes, we suspect 5-hmC could also overlap more broadly with ASD genes. To examine this possibility, we investigated 190 autism candidate genes from the Simons Foundation Autism Research Initiative (SFARI) database (as of October 2011). Even though most of the FMRP target genes and autism candidate genes are associated with synaptic functions, in contrast to FMRP target genes, we did not see 5-hmC preferentially enriched in these autism candidate genes. However, consistent with FMRP target genes, we observed that the identified DhMRs between fetus and adult were enriched more ($p<0.001$, Wilcoxon rank test; and observed to expected ratio: 1.65) in autism candidate genes relative to all RefSeq genes (Fig. 3-8A and B). The enrichment is also true when the 190 candidate genes were grouped into highly expressed genes, moderately expressed genes and weakly expressed genes (Fig. 3-8C). To further examine this correlation, we focused on the 22 syndromic genes and 21 strong candidate and suggestive genes linked to autism in SFARI database. We found that both the total counts and DhMRs of 5-hmC are enriched in some autism syndromic genes including CACN1C, SHANK3 and TSC2 (Fig. 3-9A and C) as well as strong candidate and suggestive genes including CACNA1H, ATP10A and OXTR (Fig. 3-9B and D). Importantly, DhMRs significantly overlap with strong candidate and suggestive genes (Observed-to-Expected ratio: 1.52).

Taken together, these results indicate that both stable and dynamic 5-hmC strongly associated with FMRP target genes, and 5-hmC changes are associated with autism genes. Our results suggest dysregulation of 5-hmC could be a potential contributor to the pathogenesis of neurodevelopmental disorders.

DISCUSSION

Previous studies have shown that 5-hmC is enriched in the mammalian brain (10, 21, 23), and its levels vary in different tissues (34-37). However, the genomic features of 5-hmC modification during development of human brain are still unknown. To explore this, we profiled 5-hmC at a genome-wide level in fetal and adult human brain and found that 5-hmC is increased during the development of human cerebellum, suggesting strong DNA hydroxymethylation dynamics in this brain region. In our previous study, we have discovered that 5-hmC displays dynamics during the postnatal development of mouse brain (21). Together with this finding, our current results suggest that 5-hmC-mediated epigenetic pathways might play evolutionarily conserved roles in the mammalian brain. Compared with other brain regions, the cerebellum displays a distinct pattern of both DNA methylation and gene expression (38-40). The differences between the cerebellum and other brain regions may be partially attributable to cerebellum Purkinje neurons, which are considered a primary organizer in the development of the cerebellum. Purkinje neurons exhibit a greater proportion of 5-hmC versus other types of neurons. Therefore, the cerebellum may be a specific brain region susceptible to 5-hmC changes partially through Purkinje neurons.

DNA methylation is considered a relatively stable epigenetic mark compared with histone modifications. During the last several years, DNA methylation signature were found to show certain dynamics associated with brain development and aging. Since

most approaches assessing DNA methylation levels cannot distinguish hydroxymethylation from methylation, the extent to which hydroxymethylation contributes to previously discovered methylation dynamics remains unknown. Our results suggest that, at least in the cerebellum, 5-hmC levels also change dramatically. In particular, some of the 5-hmC dynamically modified genes, such as MYOD1, have been reported to be associated with methylation dynamics during brain development and the aging process (39, 40). It is likely that 5-hmC, together with 5-mC, acts as an intermediate step for upstream regulators to regulate these gene expression changes.

5-hmC is abundant in hESCs and brain regions, and the genomic features of 5-hmC in hESCs are well characterized (28, 41). Similar to hESCs, in cerebellum, 5-hmC displayed a bimodal distribution around TSS and was enriched in gene body regions. Interestingly, compared with hESCs, we found less 5-hmC at CpG islands and CpG shores, suggesting CpG islands and shores may have a specific protective mechanism to against 5hmC modification in cerebellum. Furthermore, we find 5-hmC changes tend to occur more often outside of CpG islands (more likely to occur in CpG island shores than islands). This observation is also consistent with DNA methylation dynamics being seen most often occur outside of CpG island regions during early brain development and aging, as well as tissue and cancer differences (42, 43). For example, Numta *et al* showed that DNA methylation changes associated with developmental and aging processes in prefrontal cortex are more likely to occur outside of CpG islands (43).

Neurodevelopmental disorders, such as ASD, usually present years after birth. However, the molecular pathogenesis is thought to occur early during pregnancy or around birth. Consistent with the finding that DNA methylation changes in the brain during early life stages, we found 5-hmC also shows a positive correlation with cerebellum development both in mouse and in human. The cerebellum of the human brain plays an important role in motor control as well as motor learning. In additional to this well-established role, there has been a growing recognition that the cerebellum is also involved in cognitive functions such as attention and language, and emotional control, such as fear and pleasure responses (44, 45). Our findings in human cerebellum could have further implications for 5-hmC in the pathogenesis of neurodevelopmental disorders, which are associated with cognitive impairment, stereotypic movement, *etc*. Our earlier study revealed the global level of 5-hmC is negatively correlated with the Rett syndrome protein MeCP2 dosage, implying loss of MeCP2 could influence DNA methylation at DhMRs during brain development (21). In the present study, we have gained several new insights. We find that FMRP target genes are highly enriched with 5-hmC modifications and subject to changes during cerebellum development. This observation may be unique to FMRP target genes, since we did not see 5-hmC enrichment in the targets of other RNA binding proteins such as TDP-43, whose dysfunction can cause ALS, Amyotrophic lateral sclerosis (ALS), a neurodegenerative disease. FMRP can inhibit protein translation of target mRNAs that are involved in neuron plasticity, the balance between sensitization and desensitization responding to neuron activity. Therefore, FMRP modulated protein concentration may be critical for normal neuronal function and sensitive to gene dosage. The statistically significant overlap between 5-hmC and FMRP target genes suggests 5-

hmC may contribute to proper brain function and development via epigenetic regulation of these gene transcription activities. For example, we see a dramatic decrease of 5-hmC levels in the FMRP target gene MAP1B, which encodes a protein that belongs to the microtubule-associated protein family. MAP1B protein is involved in microtubule assembly, which is an essential step for neurogenesis. Still, the molecular mechanism remains to be determined.

We found that DhMRs between fetus and adult are associated more with genes. Moreover, we also found 5-hmC changes are statistically significantly associated with autism candidate genes, including several well-characterized autism candidate genes such as SHANK3, NLGN3 and TSC2. FMRP targets and ASD genes can be grouped into several functional categories, such as synaptic cell adhesion molecules, the NMDAR complex, the mTOR pathway, and so on, most of which are associated with synaptic functions (32, 46). Notably, synaptic dysfunction is critical to the development of autistic features and intellectual disabilities. Our observations suggest that aberrant DNA hydroxymethylation of these genes may potentially contribute to the etiology of autism and related neurodevelopmental disorders. TET1 is known to involved in neuronal activity-induced DNA demethylation and subsequent gene expression in mouse brain (9), suggesting 5-hmC dynamics may be regulated by TET1 during cerebellum development.

In summary, we profiled the genome-wide distribution of 5-hmC during the development of human brain. Our results indicate 5-hmC levels increase during cerebellum

development, which could contribute to proper brain function and development. Most importantly, we found DhMRs preferentially locate in FMRP target genes and ASD genes, implying that abnormal alteration of 5-hmC may contribute to neurodevelopmental disorders.

MATERIALS AND METHODS

Cerebellum DNA sample preparation

Fetal cerebellum DNA sample was from The State Key Laboratory of Medical Genetics (SKLMG), Hunan, China. The female and male adult cerebellum samples were from The Emory Alzheimer's Disease Research Center (ADRC). Genomic DNA was purified by Proteinase K digestion (0.667g/l Proteinase K in 10 mM Tris-HCl, pH 8.5, 5 mM EDTA, 0.2% SDS, 200mM NaCl, overnight at 55℃) followed by extraction with equal volume 25:24:1 phenol:chloroform:isoamyl alcohol saturated with 10mM Tris, pH 8.0, 1mM EDTA. Purified genomic DNA was precipitated with an equal volume of isopropanol and resuspended in 10mM Tris-HCl, pH 8.0.

5-hmC dot blot and capture

5-hmC dot-blot was performed as described previously (22). The primary antibody used is anti-5-hmC antibody (1:10,000, #39769, Active Motif). For 5-hmC capture, genomic DNA was sonicated into ~200bp size by Misonix 3000 (microtip, 4 pulses, 27s each, 1min rest, power output 2, on ice). Fragment size was verified by agarose gel electrophoresis. The following 5-hmC capture steps were performed as described previously (21).

Next generation sequencing of 5-hmC-enriched DNA

5hmC captured libraries were generated by the NEBNext ChIP-Seq Library Prep Reagent

Set for Illumina according to the manufacturer's protocol. Briefly, 25 ng of input genomic DNA or 5-hmC-captured DNA were used. DNA fragments between 150 and 300 bp were gel-purified after the adapter ligation step. PCR-amplified DNA libraries were quantified on an Agilent 2100 Bioanalyzer and diluted to 6-8 pM for subsequent cluster generation and sequencing. We performed 38-cycle single-end sequencing using Version 4 Cluster Generation and Sequencing Kits and Version 7.0 recipes. Image processing and sequence extraction were done using the standard Illumina Genome Analyzer software and pipelines developed in house at the Department of Human Genetics, Emory University.

Sequence alignment and peak identification

Human FASTQ sequence files were aligned to human (NCBI36, hg18) references using Bowtie 0.12.6 with no more than 2 mismatches within the first 25bp (47). After alignment, a custom computational pipeline was used to retain only non-duplicate unique genomic matches. 5hmC enrichment peaks were determined using a Model-based Analysis of ChIP-Seq (MACS) against genomic DNA input. Parameters used for analysis were: effective genome size = 2.7e+09; Tag size = 38; Bandwidth = 200; P-value cutoff = 1.00e-05.

DhMR identification and annotation

To identify fetus- and adult-specific DhMRs in human brain, we employed a Poisson-based peak identification algorithm (MACS) using aligned 5-hmC-enriched tags. DhMRs

were determined among all pairs of 5-hmC-enriched samples by directly comparing one sample to another in each direction. Parameters used were: effective genome size = 2.7e+09; Tag size = 38; Bandwidth = 200; P-value cutoff = 1.00e-08.

Association of DhMRs with genomic features was achieved by overlapping defined sets of DhMRs with known genomic features obtained from UCSC Tables for NCBI36/hg18: RefSeq Whole Gene, 5'UTR, Exon, Intron, 3'UTR, +/-500bp of TSS, RefSeq Intergenic, CpG Islands, and CpG Island shores (+/-2kb of CpG islands). DhMRs were assigned to a given genomic feature if overlapping $\geq$ 1bp. In order to determine the fold-change from expected values, the percentage of total DhMRs within a defined set were divided by the percentage expected to overlap each genomic feature by chance, based on the percentage of genomic space occupied by that genomic feature. All the statistical analysis and data processing were performed using R (http://www.r-project.org/).

FMRP target genes were obtained from Darnell *et al* by HITS-CLIP experiment (32). TDP-43 target genes were obtained from Polymenidou *et al* by HITS-CLIP experiment (48). Autism related genes were obtained from the SFARI database of autism candidate genes (http://gene.sfari.org). Lists of highly, moderately or weakly expressed genes were generated according to human cerebellum RNA-Seq data in BrainSpan: Atlas of the Developing Human Brain database (http://www.brainspan.org/).

Gene ontology (GO) analysis

GO analyses were performed as previously described using DAVID Bioinformatics Resources 6.7 Functional Annotation Tool (49). Gene sets were identified by joining subsets of DhMRs with RefSeq Tables obtained from the UCSC genome browser Tables.

ACCESSION NUMBER

Sequencing data have been deposited to GEO with accession number GSE40539.

ACKNOWLEDGEMENTS

We would like to thank the members of Jin lab and Viren Patel at the Department of Human Genetics for their assistance. We thank Michael Santoro, Joshua Suhl, and C. Strauss for critical reading of the manuscript.

FUNDING

This study was supported in part by the National Institutes of Health (NS051630 and MH076090 to P.J.; MH089606 and HD24064 to S.T.W.), the Emory Genetics Discovery Fund, and the Autism Speaks grant (#7660 to X.L.).

COMPETING INTERESTS

The authors declare that they have no competing interests.

Figure 3-1

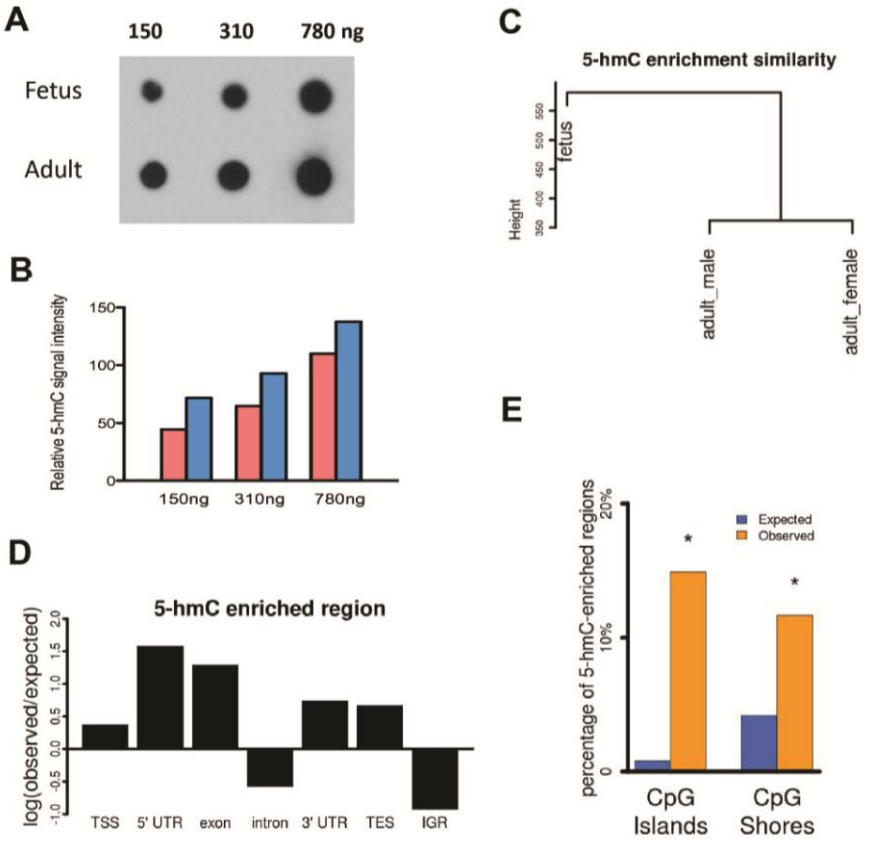

Figure 3-1. Global DNA hydroxymethylation dynamics in the developing human cerebellum. (A) 5-hmC dot blot analysis shows a significant increase of total 5-hmC levels in adult cerebellum compared to fetal cerebellum. Each column indicates the total amount of DNA used. (B) Quantitative measurement of total intensity of 5-hmC levels shown in (A). (C) Cluster dendrogram analysis using the 5-hmC-enriched regions either present in fetus or two adult samples across all samples. (D) The relative enrichment of 5-hmC distribution across distinct genomic regions. (E) 5-hmC is preferentially enriched in CpG islands and CpG island shores. Asterisk indicates statistical significance of

observed distribution compared with expected distribution ($p<0.001$, Pearson's chi-squared test).

Figure 3-2

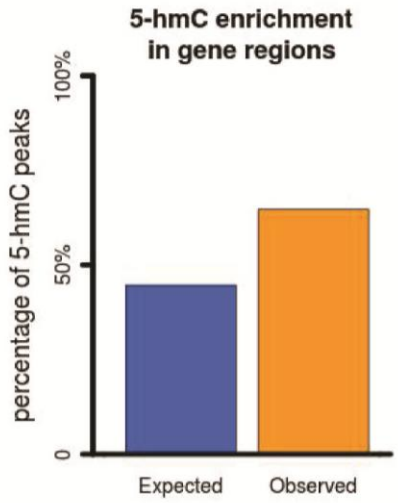

Figure 3-2. 5-hmC is strongly associated with gene regions. Shown is the observed percentage of total 5-hmC peaks compared with the expected percentage of 5-hmC peaks by random selection.

Figure 3-3

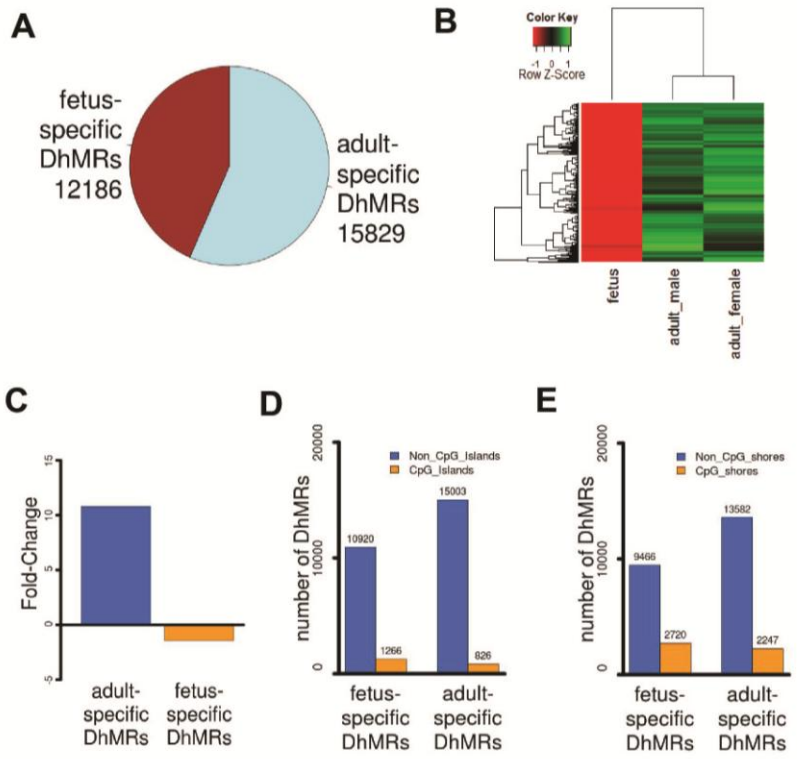

Figure 3-3. Unique genomic features of dynamic 5-hmC in fetal and adult cerebellums. (A) Total number of fetus-specific DhMRs and adult-specific DhMRs. (B) Heatmap of the top 500 DhMRs that show the most significant differences between fetus and adult. Green color represents more 5-hmC counts, red color represents less 5-hmC counts. (C) Average fold-change of adult-specific DhMRs and fetus-specific DhMRs. (D) Fetus- and adult-specific DhMRs distribution in CpG islands. (E) Fetus- and adult-specific DhMRs distribution in CpG island shores. In (D) and (E), yellow bar indicates numbers overlapping with CpG islands or shores, blue bar indicates numbers not overlapping with CpG islands or shores.

Figure 3-4

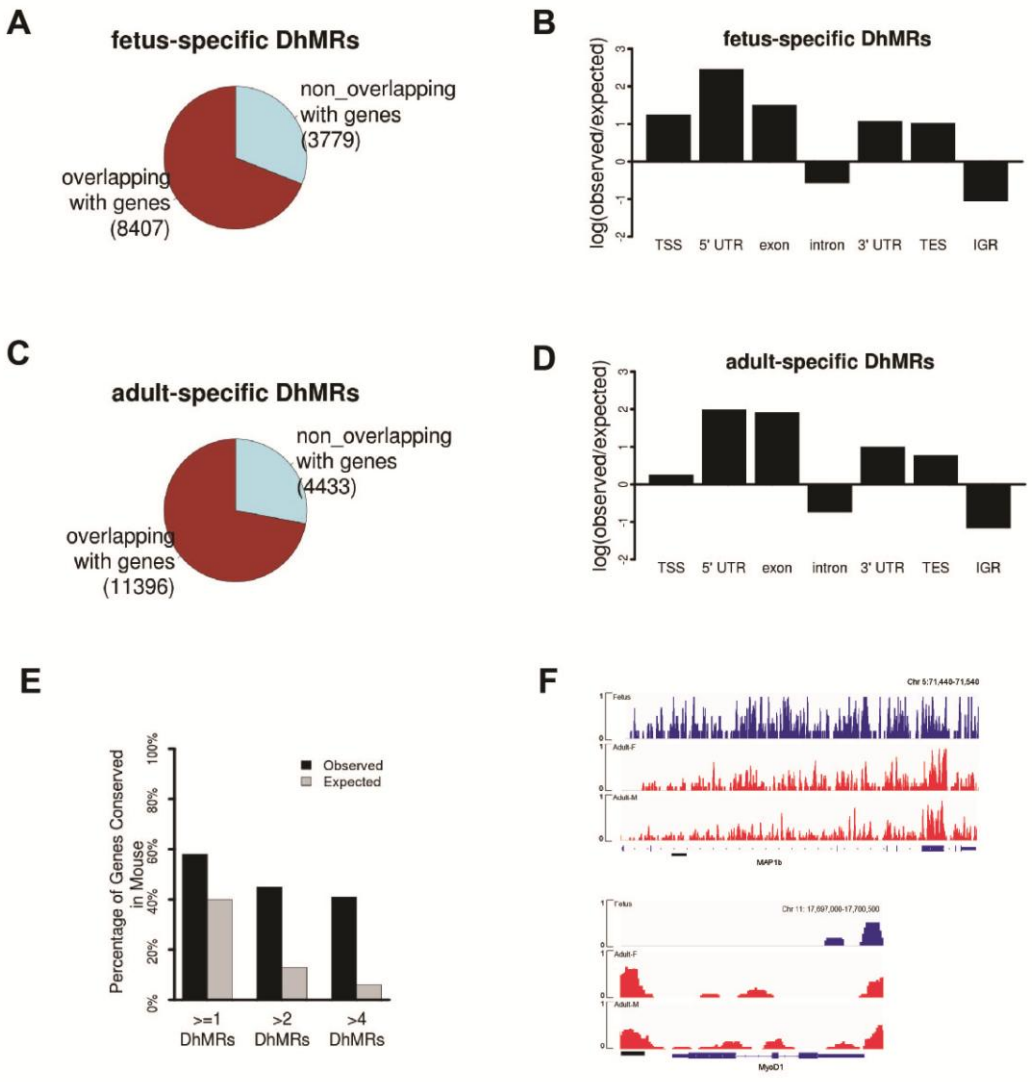

Figure 3-4. Genomic features of DhMRs during cerebellum development. (A) Number of fetus-specific DhMRs overlapping with genes. (B) The relative enrichment of fetus-specific DhMRs distribution across distinct genomic features. (C) Number of adult-specific DhMRs overlapping with genes. (D) The relative enrichment of adult-specific DhMRs distribution across distinct genomic features. (E) Genes associated with

DhMRs are highly conserved between human and mouse. Showing is observed and expected percentage of human genes associated with at least 1, 2 and 4 DhMRs that are conserved in mouse. (F) IGV Genome Browser track showing 5-hmC levels of MYOD1 and MAP1b in fetus and two adult cerebellums.

Figure 3-5

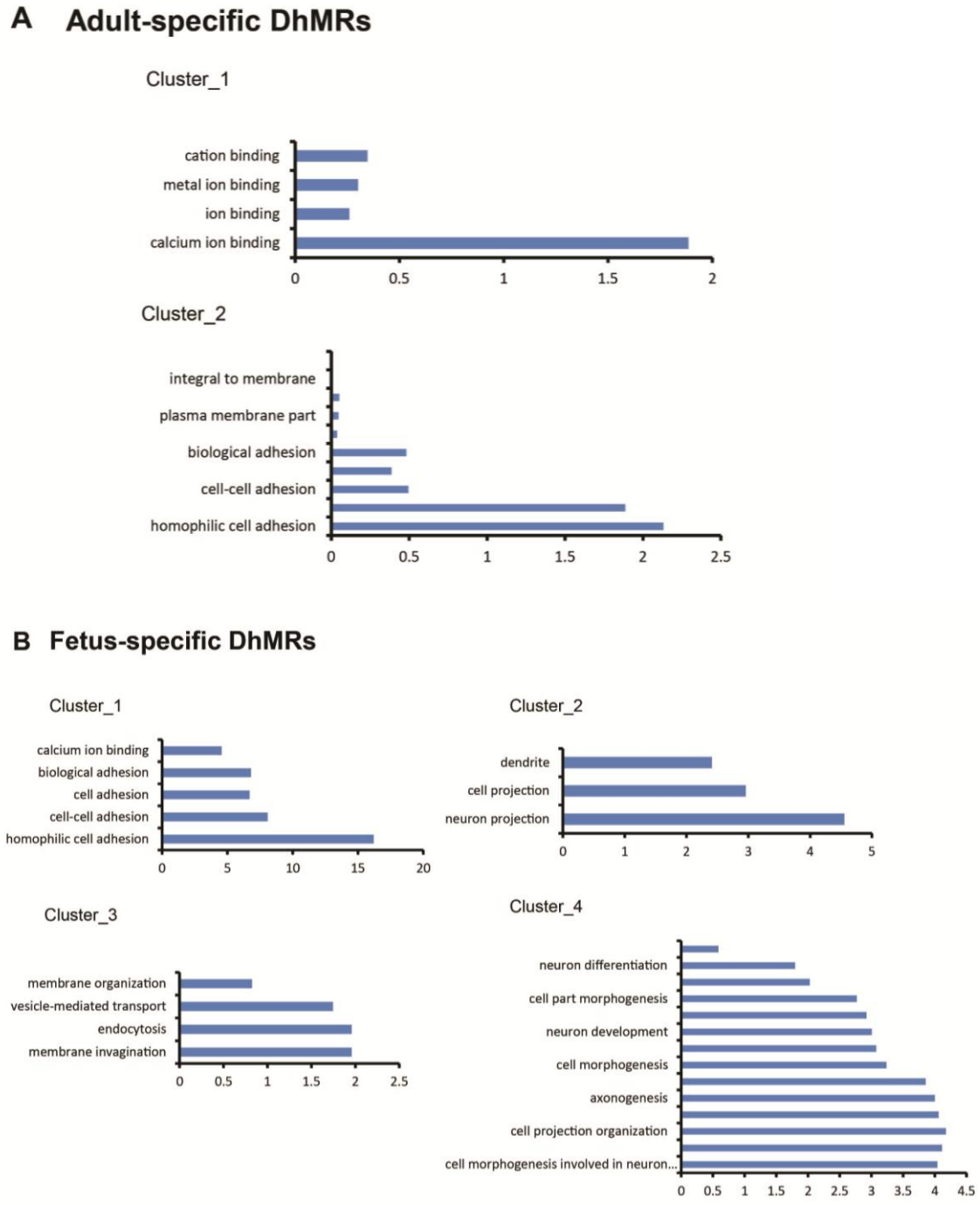
A Adult-specific DhMRs
Cluster_1
cation binding
metal ion binding
ion binding
calcium ion binding
0
0.5
1
1.5
2
Cluster_2
integral to membrane
plasma membrane part
biological adhesion
cell-cell adhesion
homophilic cell adhesion
0
0.5
1
1.5
2
2.5
B Fetus-specific DhMRs
Cluster_1
calcium ion binding
biological adhesion
cell adhesion
cell-cell adhesion
homophilic cell adhesion
0
5
10
15
20
Cluster_2
dendrite
cell projection
neuron projection
0
1
2
3
4
5
Cluster_3
membrane organization
vesicle-mediated transport
endocytosis
membrane invagination
0
0.5
1
1.5
2
2.5
Cluster_4
neuron differentiation
cell part morphogenesis
neuron development
cell morphogenesis
axonogenesis
cell projection organization
cell morphogenesis involved in neuron...
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

Figure 3-5. Gene ontology analysis reveals that cerebellum 5-hmC related development dynamics are associated with cellular functional groups. (A) The top two gene ontology clusters enriched in adult-specific DhMR associated genes in indicated gene ontology categories, X axis is denoted by -log10 (corrected p) value. (B) The top four gene ontology clusters enriched in fetus-specific DhMR associated genes in indicated gene ontology categories, X axis is denoted by -log10 (corrected p) value.

Figure 3-6

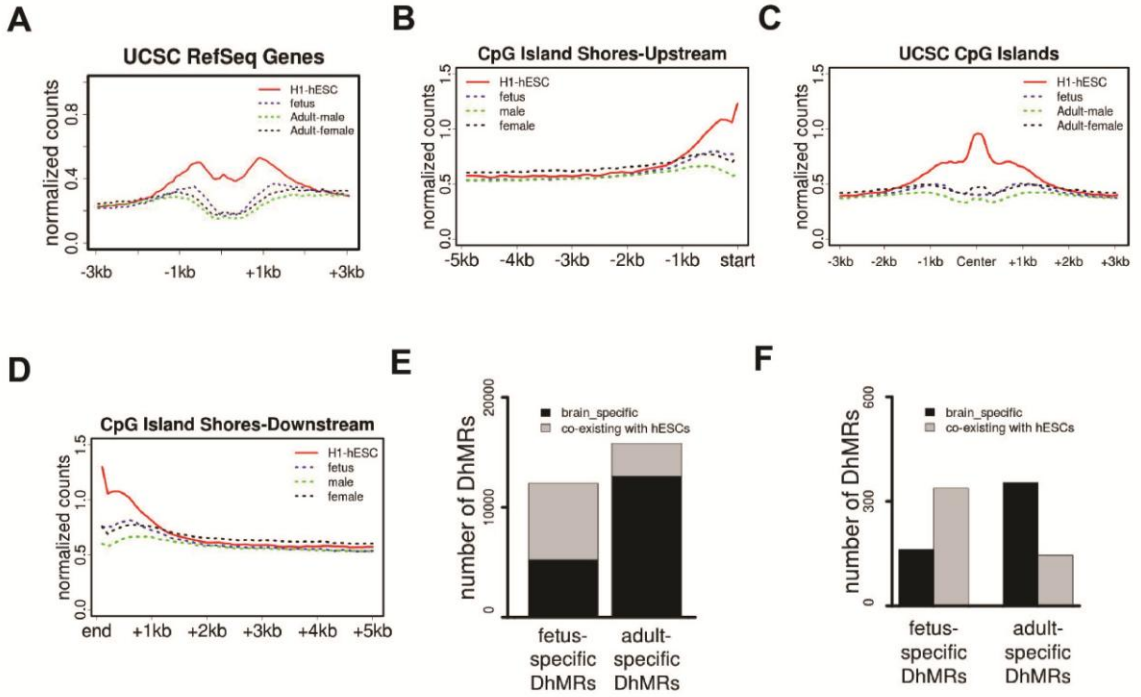

Figure 3-6. Fetus-specific DhMRs show more epigenetic memories that are present in embryonic stem cells. (A) Enrichment of 5-hmC at RefSeq gene TSS boundary regions (+/- 3kb) in H1 hESCs, fetal and adult cerebellums. (B) Enrichment of 5hmC at CpG island shores upstream boundary regions (-5kb) in H1 hESCs, fetal and adult cerebellums. (C) Enrichment of 5-hmC at CpG island boundary regions (+/- 3kb) in H1 hESCs, fetal and adult cerebellums. (D) Enrichment of 5-hmC at CpG island shores downstream boundary regions (-5kb) in H1 hESCs, fetal and adult cerebellums. (E) 5-hmC in fetus-specific DhMRs compared with H1 hESCs. Regions also showing 5-hmC enrichment in H1 hESCs are highlighted in gray. Regions showing absence of 5-hmC in

H1 hESCs are highlighted in black. (F) Top 500 most significant DhMRs by fold-change compared with H1 hESCs; grey and black color are the same as (E).

Figure 3-7

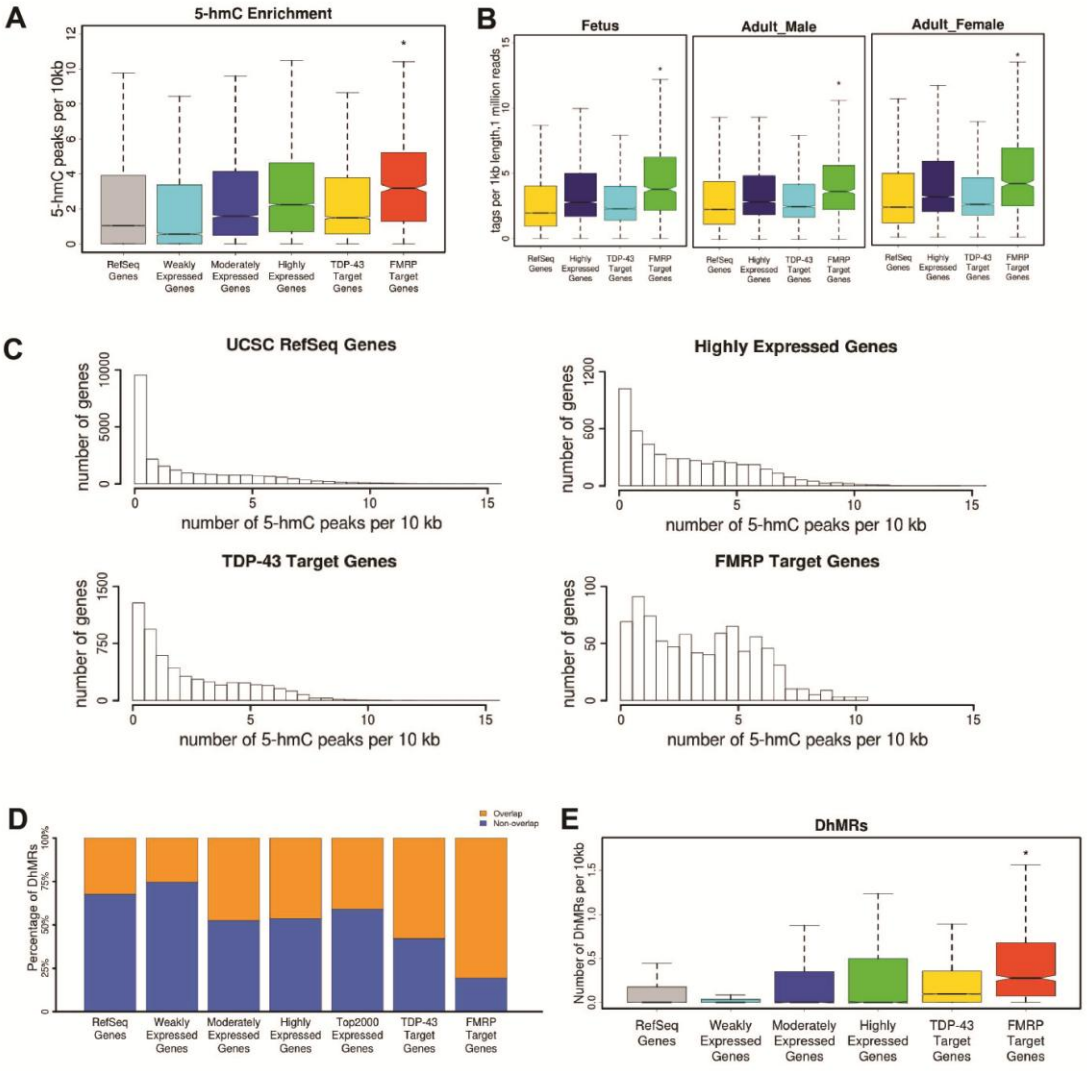

Figure 3-7. FMRP target genes have strong 5-hmC enrichment and are significantly associated with DhMRs. (A) Number of 5-hmC peaks per 10kb length in RefSeq genes, weakly, moderately, highly expressed genes, TDP-43 target genes and FMRP target genes. (B) Box plots of hydroxymethylation levels among all RefSeq genes, highly expressed genes, TDP-43 target genes and FMRP target genes in fetus (left panel) and

two adults (middle and right panels). Asterisk indicates significantly more 5-hmC levels compared with all others ($p<0.001$, Wilcoxon rank test). (C) Histograms of gene number summary by number of 5-hmC peaks per 10kb in each gene of all RefSeq genes, highly expressed genes, TDP-43 target genes and FMRP target genes. (D) The composition of RefSeq genes, weakly, moderately, and highly expressed genes, top 2000 most expressed genes, TDP-43 target genes, FMRP target genes that are associated with DhMRs between fetus and adult cerebellums. (E) Box plots of normalized DhMRs among all groups mentioned in (D).

Figure 3-8

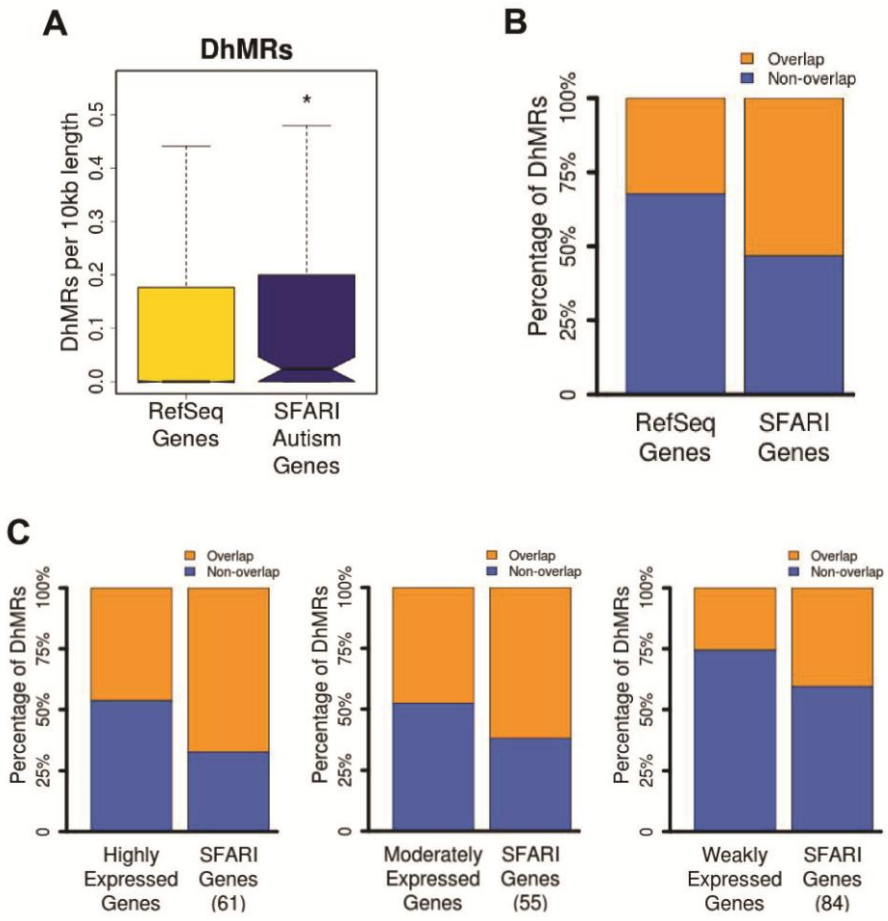

Figure 3-8. SFARI autism candidate genes are preferentially associated with developmentally dynamic hydroxymethylation regions. (A) DhMRs distribution in all RefSeq genes and SFARI autism candidate genes. Genes are normalized by 10kb length when counting DhMRs. Asterisk indicates a statistical significant difference in 5-hmC levels ($p<0.001$, Wilcoxon rank test). (B) Percentage of all RefSeq genes and SFARI autism genes that are associated with DhMRs. (C) Percentage of highly expressed genes and 61 highly expressed SFARI autism genes(left), moderately expressed genes

and 55 related SFARI autism genes (middle), weakly expressed genes and 84 related SFARI autism genes (right) that are overlapped with DhMRs.

Figure 3-9

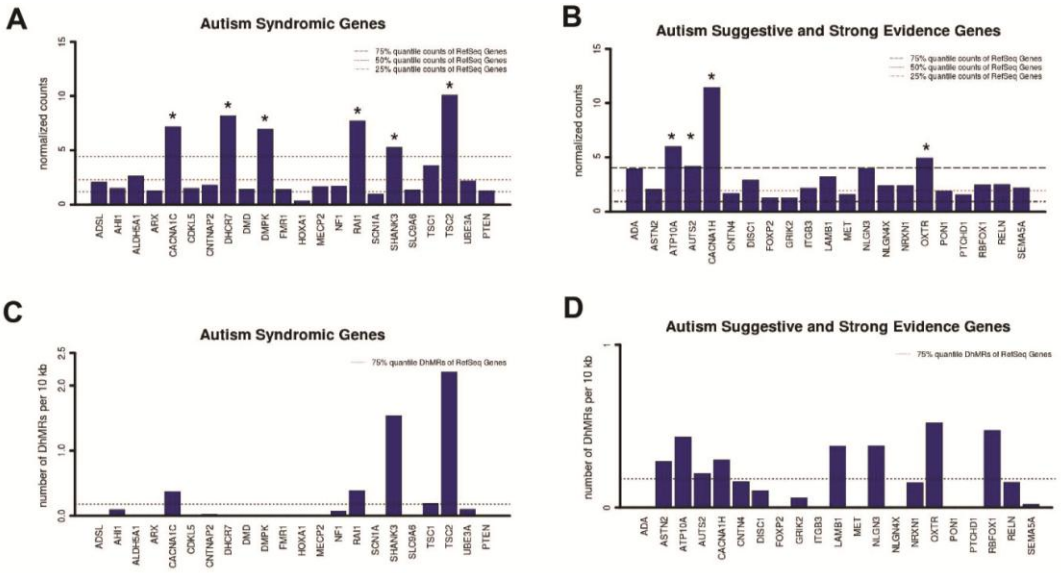

Figure 3-9. Autism suggestive and strong evidence genes are significantly associated with dynamic 5-hmCs. (A) 5-hmC level in 22 autism syndromic genes. (B) 5-hmC level in autism suggestive and strong evidence genes. Asterisk in (A) and (B) represents genes that have more 5-hmC modification compared with 75% of all genes. (C) Dynamic hydroxymethylation distribution in 22 autism syndromic genes. (D) Dynamic hydroxymethylation distribution in 21 autism suggestive and strong evidence genes.

REFERENCES

1. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33 Suppl**, 245-254.

2. Okano, M., Bell, D.W., Haber, D.A. and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247-257.

3. Li, E., Bestor, T.H. and Jaenisch, R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915-926.

4. Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., Zhang, Y. and Sun, Y.E. (2010) Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science*, **329**, 444-448.

5. Hirabayashi, Y. and Gotoh, Y. (2010) Epigenetic control of neural precursor cell fate during development. *Nat. Rev. Neurosci.*, **11**, 377-388.

6. Feng, J., Zhou, Y., Campbell, S.L., Le, T., Li, E., Sweatt, J.D., Silva, A.J. and Fan, G. (2010) Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nat. Neurosci.*, **13**, 423-430.

7. Miller, C.A., Gavin, C.F., White, J.A., Parrish, R.R., Honasoge, A., Yancey, C.R., Rivera, I.M., Rubio, M.D., Rumbaugh, G. and Sweatt, J.D. (2010) Cortical DNA methylation maintains remote memory. *Nat. Neurosci.*, **13**, 664-666.

8. Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.*, **23**, 185-188.

9. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. and Song, H. (2011) Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, **145**, 423-434.

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

11. Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L. *et al.* (2009) Conversion of 5-

methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930-935.

12. Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129-1133.

13. Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G. *et al.* (2011) Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, **8**, 200-213.

14. Quivoron, C., Couronne, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.H. *et al.* (2011) TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. *Cancer Cell*, **20**, 25-38.

15. Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X. *et al.* (2011) Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell*, **20**, 11-24.

16. Li, Z., Cai, X., Cai, C.L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.C. and Xu, M. (2011) Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. *Blood*, **118**, 4509-4518.

17. Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R. *et al.* (2010) Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*, **468**, 839-843.

18. Ko, M., Bandukwala, H.S., An, J., Lamperti, E.D., Thompson, E.C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S.B. and Rao, A. (2011) Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proc. Natl Acad. Sci. USA*, **108**, 14566-14571.

19. Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G. *et al.* (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, **477**, 606-610.

20. Inoue, A. and Zhang, Y. (2011) Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science*, **334**, 194.

21. Szulwach, K.E., Li, X., Li, Y., Song, C.X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I. *et al.* (2011) 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.*, **14**, 1607-1616.

22. Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X. *et al.* (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.*, **29**, 68-72.

23. Jin, S.G., Wu, X., Li, A.X. and Pfeifer, G.P. (2011) Genomic mapping of 5-hydroxymethylcytosine in the human brain. *Nucleic Acids Res.*, **39**, 5015-5024.

24. Feng, S., Jacobsen, S.E. and Reik, W. (2010) Epigenetic reprogramming in plant and animal development. *Science*, **330**, 622-627.

25. Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425-432.

26. Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J. and Helin, K. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*.

27. Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P. *et al.* (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, **473**, 394-397.

28. Szulwach, K.E., Li, X., Li, Y., Song, C.X., Han, J.W., Kim, S., Namburi, S., Hermetz, K., Kim, J.J., Rudd, M.K. *et al.* (2011) Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet.*, **7**, e1002154.

29. Wang, T., Bray, S.M. and Warren, S.T. (2012) New perspectives on the biology of fragile X syndrome. *Curr. Opin. Genet. Dev.*, **22**, 256-263.

30. Santoro, M.R., Bray, S.M. and Warren, S.T. (2012) Molecular mechanisms of fragile X syndrome: a twenty-year perspective. *Annu. Rev. Pathol.*, **7**, 219-245.

31. Tan, L. and Shi, Y.G. (2012) Tet family proteins and 5-hydroxymethylcytosine in development and disease. *Development*, **139**, 1895-1902.

32. Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W. *et al.* (2011) FMRP stalls ribosomal

translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247-261.

33. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A. *et al.* (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron*, **74**, 285-299.

34. Li, W. and Liu, M. (2011) Distribution of 5-hydroxymethylcytosine in different human tissues. *J. Nucleic Acids*, **2011**, 870726.

35. Ruzov, A., Tsenkina, Y., Serio, A., Dudnakova, T., Fletcher, J., Bai, Y., Chebotareva, T., Pells, S., Hannoun, Z., Sullivan, G. *et al.* (2011) Lineage-specific distribution of high levels of genomic 5-hydroxymethylcytosine in mammalian development. *Cell Res.*, **21**, 1332-1342.

36. Kinney, S.M., Chin, H.G., Vaisvila, R., Bitinaite, J., Zheng, Y., Esteve, P.O., Feng, S., Stroud, H., Jacobsen, S.E. and Pradhan, S. (2011) Tissue-specific distribution and dynamic changes of 5-hydroxymethylcytosine in mammalian genomes. *J. Biol. Chem.*, **286**, 24685-24693.

37. Nestor, C.E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J.M., Harrison, D.J. and Meehan, R.R. (2012) Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res.*, **22**, 467-477.

38. Ladd-Acosta, C., Pevsner, J., Sabunciyan, S., Yolken, R.H., Webster, M.J., Dinkins, T., Callinan, P.A., Fan, J.B., Potash, J.B. and Feinberg, A.P. (2007) DNA methylation signatures within the human brain. *Am. J. Hum. Genet.*, **81**, 1304-1315.

39. Hernandez, D.G., Nalls, M.A., Gibbs, J.R., Arepalli, S., van der Brug, M., Chong, S., Moore, M., Longo, D.L., Cookson, M.R., Traynor, B.J. *et al.* (2011) Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum. Mol. Genet.*, **20**, 1164-1172.

40. Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R. *et al.* (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.*, **5**, e1000602.

41. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. and Jacobsen, S.E. (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.*, **12**, R54.

42. Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178-186.

43. Numata, S., Ye, T., Hyde, T.M., Guitart-Navarro, X., Tao, R., Wininger, M., Colantuoni, C., Weinberger, D.R., Kleinman, J.E. and Lipska, B.K. (2012) DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.*, **90**, 260-272.

44. Schmahmann, J.D. and Caplan, D. (2006) Cognition, emotion and the cerebellum. *Brain*, **129**, 290-292.

45. Schmahmann, J.D. (2010) The role of the cerebellum in cognition and emotion: personal reflections since 1982 on the dysmetria of thought hypothesis, and its historical evolution from theory to therapy. *Neuropsychol. Rev.*, **20**, 236-260.

46. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368-372.

47. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

48. Polymenidou, M., Lagier-Tourenne, C., Hutt, K.R., Huelga, S.C., Moran, J., Liang, T.Y., Ling, S.C., Sun, E., Wancewicz, E., Mazur, C. *et al.* (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci.*, **14**, 459-468.

49. Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44-57.

Chapter 4

Discussion and Future Directions

Potential distinct roles for TET1 and TET2

Our study suggests that the significant increase of 5hmC during reprogramming is mainly owing to activation of the TET1 protein in human iPSCs, in contrast to previous observations that both Tet1 and Tet2 are upregulated in mouse iPSCs. This raises the important question of: whether TET1 and TET2 functions are redundant in ES cells. Tet2 but not Tet1 has been implicated in haematopoiesis and human TET2 mutations are seen frequently in various leukemias including chronic myelomonocytic leukemia (CMML), acute myeloid leukemias (AML) and so on (Ko et al., 2010). Compared with TET1, TET2 is lacking a CXXC DNA binding domain, which is suspected to cause a loose chromatin association.

As discussed in the second chapter of this thesis, mouse ESCs are different from hESCs in terms of signaling maintenance of pluripotency, and their epigenetic properties such as X-chromosome inactivation status in female lines. Human ES pluripotency (primed pluripotency) depends mainly on FGF and Activin-Nodal signaling pathways, whereas mouse pluripotency (naive/ground-state pluripotency) depends on LIF-STAT pathways. Thus, it is possible that TET1 and TET2 have distinct roles in regulating different pluripotency states, with TET2 being involved in naive pluripotency and TET1 functioning in primed pluripotency. It is also possible that TET1-mediated 5hmC modification is unique in humans regardless of different pluripotent stages. As TET1 and TET2 are dispensable for maintaining the pluripotency of stem cells, and their loss is compatible with embryonic and postnatal development, it is likely that TET2 expression is not under positive selection for stem cell functions during evolution, and is thus eventually silenced in human pluripotent stages. Based on preliminary data, TET2 is

also highly expressed in mouse primed pluripotent stem cells. As human iPS cells are equivalent to primed pluripotent state, TET1 is very likely to be the sole enzyme in the different human pluripotent stages.

Quite interestingly, in one study, TET2 and TET3 were associated both *in vitro* and *in vivo* with O-linked β-N-acetylglucosamine (O-GlcNAc) transferase (OGT), an enzyme that catalyses the addition of O-GlcNAc onto serine and threonine residues through their hydroxyl groups (Chen et al., 2013). OGT is highly evolutionarily conserved, with 85% similarity between *Caenorhabditis elegans* and humans. O-GlcNAc-modified proteins are almost ubiquitous in the cytoplasm and nucleus and are abundant in all eukaryotic cells. A broad range of different chromatin proteins and transcription factors, some of which are physically associated with chromatin, are O-GlcNAc-modified. For example, SP1, MYC, p53, OCT4 and RNA polymerase II transcription factors contain O-GlcNAc. Moreover, TET2 and TET3 are found to promote OGT activity, and TET2 and TET3 are important for the chromatin association of OGT (Chen et al., 2013). Depletion of TET2 from ES cells prevents the association of OGT with chromatin. TET2–OGT interaction is important for the modification of histone H2B at Ser112 by O-GlcNAc. In contrast, OGT does not influence hmC activity, although it can add hydroxyl groups on proteins.

Because the OGT expression level is higher in stem cells than in fibroblasts, its function in stem cells is very likely important. O-GlcNAcylation directly regulates the pluripotency network. Blocking O-GlcNAcylation disrupts mouse ES cell self-renewal, and affects the reprogramming of somatic cells to iPSCs. The reprogramming factors OCT4 and SOX2 are O-GlcNAcylated in pluripotent stages, and the modification is

rapidly removed in somatic stages (Jang et al., 2012). TET2 expression is missing in human pluripotent stem cells, thus we don't know whether TET2-mediated-OGT regulation is critical for stem cell functions, or whether the regulation is important for downstream functions. In a different study, Tet1 and Tet2 were found to be required for the binding of Ogt to chromatin, thus affecting Tet1 activity (Vella et al., 2013). An explanation for the discrepancy between different studies of TET1 and OGT interaction remains elusive.

DNA methylation and demethylation

Our study demonstrates that TET1-mediated-hydroxymethylation is important for reprogramming. As suggested, one important implication is that this modification is critical for DNA demethylation. Because DNA demethylation can be DNA replication dependent or DNA replication independent, thus, it is still unknown whether the TET1-mediated hydroxymethylation participates in the active DNA demethylation or passive DNA demethylation, or both. Maintenance of DNA methylation patterns requires the DNA methyltransferase DNMT1 and its binding partner UHRF1 (Law and Jacobsen, 2010). UHRF1 shows a more than 10-fold greater binding affinity for hemi-5mC than hemi-5mC binding *in vitro* (Hashimoto et al., 2012). In addition, the binding activity of recombinant DNMT1 is reduced more than 60-fold for hemi-5hmC. These data imply that the TET-mediated hydroxymethylation of a methylated CG site can block maintenance of methylation during cell division and eliminate 5mC in a replication-dependent or passive manner. Because reprogramming depends on cell division, we

cannot exclude the possibility that 5hmC may function via passive DNA demethylation. On the other hand, studies imply that active demethylation can contribute to reprogramming. Fusion ESCs or EGCs with differentiated cells such as lymphocytes or fibroblasts, results in heterokaryon (contains two haploid nuclei) formation. During this cell-fusion-mediated reprogramming, the somatic epigenome is reset and genome methylation is modified. As a result, gene expression of the differentiated cell is reset to a state that resembles the pluripotent stage (Piccolo et al., 2013). Tet2 is important for the rapid activation of pluripotency-associated genes induced after fusion with EGCs and it oxidizes the 5mC at the somatic OCT4 locus. Because there is no cell division for heterokaryon induction, the cell fusion study reveals that 5hmC can contribute to active DNA demethylation.

Future directions for hydroxymethylation studies

Based on a single base resolution hydroxymethylation analysis by TAB-Seq, the pattern of hydroxymethylation is mosaic, which is similar to 5mC modification pattern. Unlike 5mC CG-methylation, the 5hmC modification level is around 0-30% in a given locus, which is similar to non-CG methylation in embryonic stem cells (Lister et al., 2009). Thus, it is interesting to know the mechanism by which a given CpG in rest cells is not hydroxymethylated. One possible mechanism is that 5hmC modification is cell cycle dependent, meaning it is enriched in a particular cycle phase. Nevertheless, how this 0-30% hydroxymethylation in a given locus contributes to gene function remains unknown.

Another important and unanswered question is how 5hmC maintains its fidelity during cell replication (Fig. 4-1). Once established, 5hmC must be stably maintained. Because TET family proteins cannot direct catalyze 5C to 5hmC, during each cell division, TET proteins must function through the 5mC intermediate to modify newly added cytosine. In mammals, DNA methylation patterns are established by the *de novo* methyltransferases DNMT3a and DNMT3b and maintained by DNMT1. During cell division, DNMT1 is associated with replication foci and functions to restore hemimethylated DNA caused by DNA replication to the fully methylated state (Chuang et al., 1997). In contrast, non-CG methylation maintenance in mammals has not been well characterized. In plants, however, non-CG methylation occurs frequently and the mechanism is well studied. Evidence suggests that SUVH4, 5 and 6, which catalyses histone 3 lysine 9 dimethylation (H3K9me2) modifications, are required for the maintenance of CHG (H=A,T,C) methylation (Jackson et al., 2002; Malagnac et al., 2002). Because the characteristics of non-CG methylation and 5hmC are very similar, understanding how non-CG methylation is maintained will shed light on the maintenance mechanism of 5hmC.

Figure 4-1

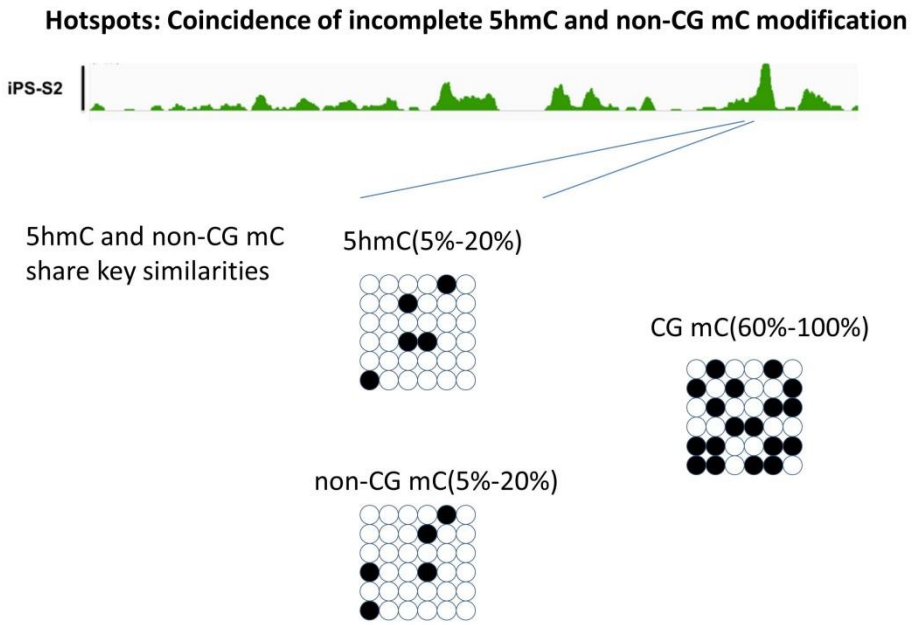

Figure 4-1. Similarity between non-CG mC and 5hmC modification. How non-CG mC maintains fidelity after cell division? How 5hmC maintains fidelity after cell division? Do they share common mechanisms?

REFERENCES

Chen, Q., Chen, Y., Bian, C., Fujiki, R., and Yu, X. (2013). TET2 promotes histone O-GlcNAcylation during gene transcription. Nature 493, 561-564.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science 277, 1996-2000.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Res 40, 4841-4849.

Jackson, J.P., Lindroth, A.M., Cao, X., and Jacobsen, S.E. (2002). Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature 416, 556-560.

Jang, H., Kim, T.W., Yoon, S., Choi, S.Y., Kang, T.W., Kim, S.Y., Kwon, Y.W., Cho, E.J., and Youn, H.D. (2012). O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. Cell stem cell 11, 62-74.

Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R., *et al.* (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature 468, 839-843.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11, 204-220.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., *et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315-322.

Malagnac, F., Bartee, L., and Bender, J. (2002). An Arabidopsis SET domain protein required for maintenance but not establishment of DNA methylation. EMBO J 21, 6842-6852.

Piccolo, F.M., Bagci, H., Brown, K.E., Landeira, D., Soza-Ried, J., Feytout, A., Mooijman, D., Hajkova, P., Leitch, H.G., Tada, T., *et al.* (2013). Different roles for Tet1 and Tet2 proteins in reprogramming-mediated erasure of imprints induced by EGC fusion. Mol Cell 49, 1023-1033.

Vella, P., Scelfo, A., Jammula, S., Chiacchiera, F., Williams, K., Cuomo, A., Roberto, A., Christensen, J., Bonaldi, T., Helin, K., *et al.* (2013). Tet proteins connect the O-linked N-acetylglucosamine transferase Ogt to chromatin in embryonic stem cells. Molecular cell 49, 645-656.

VITA

Tao Wang was born in Lu'an, Anhui, China in 1983. He graduated from Zhenjiang No.1 High School in Zhenjiang, Jiangsu Province, 2001. He then attended at the University of Science and Technology of China. He joined GMB program at Emory University in 2008; during this period, he also obtained a Master degree in Computer Science.