Title
Epigenetic Gene Regulation in Stem Cell Differentiation and Reprogramming

Permalink
https://escholarship.org/uc/item/1ng3k215

Author
Huang, Kevin

Publication Date
2013

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA

Los Angeles

Epigenetic Gene Regulation in Stem Cell Differentiation and Reprogramming

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

Kevin Huang

2013
The regulatory capacities of epigenetic mechanisms including DNA methylation, histone modifications, and non-coding RNAs, have seen a rising interest in recent years. These epigenetic marks are pervasive and non-randomly distributed across the genome, raising intriguing questions on how epigenetics contributes to genomic features that define cellular identity and function. Unlike fixed genetic information that is shared between all cell types, epigenetics involve multiple layers of regulation and can vary dramatically across different cell types and genomic contexts. Thus, much more effort is required to procure a complete perspective of the manifold epigenetic landscape. The body of work in this dissertation focuses on epigenetic studies in the mammalian pluripotent stem cell model system. We utilize high-throughput technologies such as microarrays and next-generation sequencing (NGS) as well as leverage existing epigenetic maps to address a wide range of molecular questions on a comprehensive global scale. This dissertation is organized into three overarching themes:
First, we employed genome-wide gene expression and DNA methylation profiling tools to determine whether different cell types display unique biomarkers that can be used to distinguish them from other cell types (Chapters 2-5). We found that human embryonic stem cells (hESCs) and induced pluripotent stem cells (hiPSCs) carry distinct features in both gene expression and DNA methylation patterns, arguing in favor of the idea that these two pluripotent cell types are different. Furthermore, we compared pluripotent stem cell derived retinal pigmented epithelium (hESC-RPE and hiPSC-RPE) with fetal and adult RPE and found that all RPE cells share a core set of signature genes that distinguishes them from all other cell types. We propose these signature genes will be useful for evaluating the quality of stem-cell derived RPE. Finally, novel corneal endothelial cells (CECs) biomarkers were identified through comparing 12 other tissue types, paving the way for future studies to evaluate properties of stem-cell derived CECs.

We next examined how molecular features of pluripotent stem cells are altered during the differentiation process of stem cells (Chapters 6-8). Using the RPE differentiation paradigm, we profiled both microRNA and DNA methylation patterns in intermediate stages between pluripotent stem cells and mature RPE. These two separate studies identified subsets of dynamically regulated epigenetic marks, some of which are associated with RPE signature gene expression. Furthermore, we used a highly innovative and powerful single-cell RNA-sequencing approach to profile transcriptional changes in the early embryo beginning from mature oocyte to morula stages. This study identified a conserved genetic program describing a highly dynamic transcriptional architecture during early embryogenesis.

Finally, we took a focused analysis on how DNA methyltransferases contribute to shaping the pluripotent stem cell epigenome (Chapters 9-10). Using mouse ESCs null of DNA
methylation, we determined DNA methylation regulates a large set of genes through action with H3K27me3. Furthermore, we determined shared and unique genomic targets of each DNA methyltransferase, including novel de novo methylation activity for Dnmt1 in vivo. In the human model system, we generated iPSCs from ICF Syndrome patient fibroblasts which carry double heterozygous mutations in DNMT3B. We found DNMT3B is involved in a wave of de novo methylation during the reprogramming process and has unique genomic targets.
This dissertation of Kevin Huang is approved.

Douglas L. Black

Xinshu Grace Xiao

Amander T. Clark

Matteo Pellegrini

Guoping Fan, Chair

University of California, Los Angeles

2013
This dissertation is dedicated to my grandma, Wu Jinzhuan (伍金转).

I wish her good health, peace of mind, and all things well.
Table of Contents

Abstract of the Dissertation ........................................................................................................... ii

Tables of Contents .......................................................................................................................... vii

List of Figures ................................................................................................................................... ix

List of Tables ................................................................................................................................... xi

Acknowledgements .......................................................................................................................... xii

Vita .................................................................................................................................................. xvi

**Chapter 1: Introduction** ............................................................................................................ 1

References ....................................................................................................................................... 16

**Chapter 2: A panel of CG methylation site distinguishes human embryonic stem cells and induced pluripotent stem cells** .................................................................................................................. 28

References ....................................................................................................................................... 50

**Chapter 3: Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells** ........................................................................................................................................ 56

References ....................................................................................................................................... 68

**Chapter 4: Molecular Signatures of Primary Retinal Pigment Epithelium and Stem-Cell derived RPE cells** ..................................................................................................................................... 70

References ....................................................................................................................................... 80

**Chapter 5: Identification of Novel Molecular Markers through Transcriptomic Analysis in Human Fetal and Adult Corneal Endothelial Cells** ............................................................................................................ 81

References ....................................................................................................................................... 89
Chapter 6: Identification of miRNA signatures during the differentiation of hESCs into retinal pigment epithelial cells ................................................................. 91
References 99

Chapter 7: Analysis of DNA methylation and its impact on RNA transcription during the differentiation of human pluripotent stem cells into retinal pigment epithelial cells .................................................................................................................. 101
References 118

Chapter 8: Genetic programs in human and mouse early embryos revealed by single-cell RNA-sequencing ........................................................................... 135
References 155

Chapter 9: DNMT3B mutations leads to select demethylation and altered gene expression associated with ICF Syndrome ........................................................................... 163
References 179

Chapter 10: Gene body demethylation and H3K27me3 redistribution contribute to transcriptome deregulation in mutant embryonic stem cells null of DNA methylation .............................................................................................................. 190
References 207

Chapter 11: Concluding Remarks ................................................................................................................................. 218
References 223
<table>
<thead>
<tr>
<th>List of Figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2-1.</td>
</tr>
<tr>
<td>Figure 2-2</td>
</tr>
<tr>
<td>Figure 2-3.</td>
</tr>
<tr>
<td>Figure 3-1.</td>
</tr>
<tr>
<td>Figure 3-2.</td>
</tr>
<tr>
<td>Figure 3-3.</td>
</tr>
<tr>
<td>Figure 3-4.</td>
</tr>
<tr>
<td>Figure 3-5.</td>
</tr>
<tr>
<td>Figure 3-6.</td>
</tr>
<tr>
<td>Figure 3-7.</td>
</tr>
<tr>
<td>Figure 3-8.</td>
</tr>
<tr>
<td>Figure 4-1.</td>
</tr>
<tr>
<td>Figure 4-2.</td>
</tr>
<tr>
<td>Figure 4-3.</td>
</tr>
<tr>
<td>Figure 4-4.</td>
</tr>
<tr>
<td>Figure 4-5.</td>
</tr>
<tr>
<td>Figure 4-6.</td>
</tr>
<tr>
<td>Figure 5-1.</td>
</tr>
<tr>
<td>Figure 5-2.</td>
</tr>
<tr>
<td>Figure 5-3.</td>
</tr>
<tr>
<td>Figure 5-4.</td>
</tr>
<tr>
<td>Figure 5-5.</td>
</tr>
<tr>
<td>Figure 6-1.</td>
</tr>
<tr>
<td>Figure 6-2.</td>
</tr>
<tr>
<td>Figure 6-3.</td>
</tr>
<tr>
<td>Figure 6-4.</td>
</tr>
<tr>
<td>Figure 7-1.</td>
</tr>
<tr>
<td>Figure 7-2.</td>
</tr>
<tr>
<td>Figure 7-3.</td>
</tr>
<tr>
<td>Figure 7-4.</td>
</tr>
<tr>
<td>Figure 7-5.</td>
</tr>
<tr>
<td>Figure 7-6.</td>
</tr>
<tr>
<td>Figure 7-7</td>
</tr>
<tr>
<td>Figure 8-1.</td>
</tr>
<tr>
<td>Figure 8-2.</td>
</tr>
<tr>
<td>Figure 8-3.</td>
</tr>
<tr>
<td>Figure 8-4.</td>
</tr>
<tr>
<td>Figure 9-1.</td>
</tr>
<tr>
<td>Figure 9-2.</td>
</tr>
<tr>
<td>Figure 9-3.</td>
</tr>
<tr>
<td>Figure 9-4.</td>
</tr>
<tr>
<td>Figure 10-1.</td>
</tr>
<tr>
<td>Figure 10-2.</td>
</tr>
<tr>
<td>Figure 10-3.</td>
</tr>
<tr>
<td>Figure 10-4.</td>
</tr>
<tr>
<td>Table</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>2-1</td>
</tr>
<tr>
<td>3-1</td>
</tr>
<tr>
<td>3-2</td>
</tr>
<tr>
<td>5-1</td>
</tr>
<tr>
<td>5-2</td>
</tr>
<tr>
<td>6-1</td>
</tr>
<tr>
<td>6-2</td>
</tr>
<tr>
<td>6-3</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost, I would like to thank my two graduate advisors, Guoping Fan and Matteo Pellegrini, whose support and guidance have made all my achievements possible. Guoping has opened countless opportunities for me to study gene regulation in various model systems, and Matteo has mentored me in bioinformatics every step of the way. Their complementary expertise has taught me in-depth the limitations of both “wet” and “dry” bench, and that a crucial understanding of both fields will foster the best experimental designs. Interestingly, in almost every way Guoping’s and Matteo’s mentorship styles lie on opposite ends, though they share common goals in rigorous scientific pursuit and in my well-being. Their unique influences have shaped my perspectives of how science should be—can be—accomplished. I am forever grateful to these two professors who have taught me so much in both the sciences and in life.

Many members of the Fan and Pellegrini labs past and present have also had great impact on my graduate studies. Without their helpful discussions and assistance over the years, my graduate career would have been much less successful. In addition, I am thankful for my fellow peers who have journeyed through graduate school together with me. Their perspectives and encouragement have kept me motivated throughout difficult challenges. All together, these people have constantly reminded me that academia is not entirely about science but also forging lasting friendships.

Finally, my family and friends have given me unwavering support from the very beginning; their affection has meant more to me than I can eloquently express on paper. My parents, Bingguang Huang (黄炳光) and Xuexia Zhen Huang (甄雪霞), deserve a special mention in this dissertation because their story of sacrifice, selflessness, and hard work are a continuous source of inspiration for me. Despite growing up in an underprivileged immigrant
family, my parents have shown me boundless generosity. I will probably never be able to repay all that my parents have given me, so instead I will strive to share their spirit of generosity, cooperation, and optimism to all those I meet.

Chapter 2-5 is a collection of my work on identifying biomarkers in pluripotent stem cells and their differentiated products. These studies include:


iv) A version of the manuscript “Chen Y, Huang K, Xue Z, Nakatsu M, Deng SX, Fan G. (2013). Identification of Novel Molecular Markers through Transcriptomic Analysis in Human Fetal and Adult Corneal Endothelial Cells. Hum Mol Genet. 22(7):1271-9.” Permission to reproduce this work has been granted by Oxford University Press.
Chapter 6-8 is a collection of my work that focuses on following molecular changes through progressive developmental states. These studies include:


Chapter 9-10 describes my work on understanding the role of DNA methyltransferases in regulation pluripotency. Work in this chapter would not be possible without members of the Fan lab who have been instrumental in establishing these unique genetic models. Juehua Yu developed mouse TKO re-constituted Dnmt ESCs. Zhenshan Liu generated bisulfite-sequencing libraries. Kee-pyo Kim, Jerry Wu, and Zhigang Xue generated ICF iPSCs. Together, they have contributed to the following studies:

ii) A version of the manuscript “Huang K, Yu J, Le T, Pellegrini M, Fan G. (2013) Gene body methylation and H3K27me3 re-distribution contributes to transcriptome deregulation in mouse ESCs null of DNA methylation.”
Vita

2005-2008  B.S., Biochemistry and Cell Biology, cum laude
           Phi Beta Kappa
           University of California, San Diego
           La Jolla, California

2007      Chancellor’s Research Scholarship
           University of California, San Diego

2007      S-STEM Scholarship
           University of California, San Diego

2007      Ronald E. McNair Scholar
           University of California, San Diego

2008 – 2013 Graduate Student
           Molecular Biology Interdepartmental PhD Program
           University of California, Los Angeles
           Los Angeles, California

2009-2011 Genetics Training Grant
           University of California, Los Angeles

2011-2013 California Institute of Regenerative Medicine
           Pre-doctoral Training Grant
           University of California, Los Angeles

2011      ISSCR Travel Scholarship
           California Institute of Regenerative Medicine

2012      Molecular Biology IDP Dissertation Year Award
           University of California, Los Angeles

Publications and presentations


Chapter 1
Overview of Pluripotent Stem Cells

Pluripotent stem cells have the potential to become any cell type in the adult body and can, in theory, multiply forever in a petri dish. By contrast, differentiated cells (such as somatic cells) have less access to this potential and are fated to generate only a subset of specialized cells. Remarkably, the restricted potential in differentiated cells can be reverted back to a pluripotent-like state in a process termed cellular reprogramming. For these reasons, stem cells have been touted as a self-sustaining renewable source for regenerative medicine. However, the clinical applications of stem cells are still in its infancy due to limited data on the safety and efficacy of stem cell therapy. The efforts are further stymied by an incomplete understanding of the molecular features that endow pluripotent stem cells with their unique properties and uncertainty of whether stem-cell derived products fully resemble human tissue.

Until recently, only rare populations of cells have been considered pluripotent. The best known of these are embryonic stem cells (ESCs), derived from the inner cell mass (ICM) of an early embryo during the blastocyst stage\(^1,2\). ESCs are considered the gold standard or reference point for pluripotency, but cells must pass stringent experimental assays such as—tetraploid complementation, chimera generation, and teratoma formation\(^3\)—to unequivocally demonstrate pluripotency. In the past, pluripotent cells have also been produced using other methods. Embryonic carcinoma cells (ECCs) are pluripotent cells derived from teratocarinosas\(^4\), but these cells tend to carry genomic aberrations and therefore not ideal for genetic studies. An unfertilized egg can be coaxed to begin embryogenesis and generate parthenogenetic embryonic stem cells (pESCs)\(^5,6\). Alternatively, a mature oocyte can be enucleated and replaced with a nucleus from an unrelated cell in a process called somatic cell nuclear transfer (SCNT)\(^3\). As exemplified in the famous “Dolly” the sheep experiment, genetic material from the donor cell carry the capacity to
recapitulate embryogenesis. The process whereby one cell type is altered to a different cell type is considered a form of cellular reprogramming and Dolly demonstrated that a differentiated cell can be reprogrammed to a primitive cell type that subsequently gives rise to a full individual.

More recently, artificial reversion to pluripotency has been made facile by treating cells with a cocktail of transcription factors, commonly referred to as Yamanaka factors after the Japanese Nobel Prize laureate who made the discovery. These so called induced pluripotent stem cells (iPSCs) have become a valuable research tool and chief candidate for replacing ESCs due to more practical uses in medicine.

The advent of iPSCs came during a period of rapid expansion in high-throughput technologies, making possible genome-wide assessment of the pluripotent molecular networks. Currently, there are several well-studied factors critical for ESC specification, including the core transcriptional factors Oct4, Sox2, and Nanog. Transgenic null of Oct4 fail to survive post-implantation and differentiate toward the trophoblast lineage. Sox2 and Nanog are both also essential for early embryonic development. Other components of the ESC regulatory network have been identified through extensive RNAi screens. These studies identified several novel regulators of ESCs including Ronin, Tcl1, Tbx3, Esrrb, Chd1, Ctr9, RTF1, Cnot, Trim28 to name a few top candidates. Further investigation into the ESC pluripotency identified the downstream targets of key transcription factors through chromatin immunoprecipitation followed by whole genome array or high-throughput sequencing (ChIP-chip or ChIP-seq). Results from these sets of studies have revealed that the core transcription factors Oct4, Sox2, and Nanog are frequently co-localized to the same locus. Further analysis of ESC gene profiles helped to sub-classify the dual role of these transcription factors to promote the pluripotent state while repressing differentiation genes. Importantly, these studies have
found that Oct4, Sox2, and Nanog bound to each other’s promoters, establishing an auto-regulatory and feed-forward mechanism to propagate the pluripotency network. In addition to these core transcriptional factors, binding profiles for other transcription factors have also been mapped, providing an extraordinarily extensive map of the pluripotent state\textsuperscript{24-26}.

Recent studies have demonstrated that environmental cues can also influence the pluripotency state through a system of signaling cascades. For example, it has been shown that BMP4 and LIF affects the Smad1 and Stat3 signaling pathways\textsuperscript{26}, and ultimately on the MAP kinase pathway\textsuperscript{27,28}. Furthermore, the Wnt pathway can either promote pluripotency or repress differentiation through Tcf3 interaction with beta-catenin and groucho, respectively\textsuperscript{29,30}. It has been postulated that the dual pharmacological inhibition of the MAP kinase and Wnt pathways (so-called 2-inhibitor or 2i system) promote ESCs to a more primitive “ground” or “naïve” state. These ESCs are characterized as have greater pluripotency potential because of higher efficiency of generating chimeras and maintain two active X chromosomes in female ESCs\textsuperscript{31-34}.

Collectively, these studies show that environmental signals are critical players for affecting the pluripotency state together with transcription factors.

**Epigenetic Mechanisms**

Beyond transcription factors, many other regulatory mechanisms contribute to the complex multi-layer regulatory control in pluripotent stem cells. Epigenetics is the study of mechanisms that can alter gene expression without changing the underlying DNA sequence. Under this broad definition, many mechanisms have been considered epigenetic including DNA methylation, histone modifications, and non-coding RNA. Often times, these epigenetic mechanisms work in concert to influence both gene expression and each other. Epigenetic landscapes are extremely
complex with a vast spectrum of variations that are used to fine-tune gene expression. In order to fully understand the regulatory domains in the genome, all epigenetic regulatory forces must be considered. Recent advances in high-throughput technology have afforded the opportunity to survey epigenetic features across entire genomes, bringing forth a vibrant field of “epigenomics”-based research.

**DNA Methylation**

DNA methylation is one of the best studied epigenetic mechanisms and involves the covalent attachment of a methyl group to the 5 carbon position of cytosine. In mammals, this action is catalyzed by a family of DNA methyltransferases (Dnmts), including Dnmt1, Dnmt3a, and Dnmt3b. Loss of any of these enzymes during embryogenesis is lethal, indicating an essential role for DNA methylation during development. The prevailing hypothesis on the mechanism of action for DNA methylation involves repression via its presence on the proximal promoter. It is thought that DNA methylation suppresses gene activity either by acting as a part of a signaling pathway that recruits repressor complexes, or by sterically hindering transcription factor binding. Nevertheless, global mapping of gene promoters indicate a negative correlation between promoter methylation and gene activity, with some exceptions. Epigenomic studies using the mouse embryonic stem cells (ESCs) model revealed promoters can be sub-classified based on their CpG content. For example, proximal promoters with high density of CpG dinucleotides tend to be hypomethylated whereas promoters with low density of CpGs are hypermethylated. However, the absence of DNA methylation does not necessarily predict gene activity; many gene promoters that lack DNA methylation can also be transcriptionally inactive. Furthermore, DNA methylation patterns are different in
various cell types. We now know that different cells have their own unique DNA methylation signature and these characteristics are important for governing cell identity. For example, neural genes are repressed in non-neural tissues by promoter DNA methylation, but are unmethylated in neural cells, indicating a direct role for DNA methylation. These types of studies have revealed an immense amount about the methylation status of gene promoters in regulating gene expression and cellular differentiation.

**Gene body methylation**

Global DNA methylation mapping have revealed many novel facets of DNA methylation beyond the classical model of gene regulation. For example, outside of gene promoters, DNA methylation appears to be highly enriched within the gene body (the transcribed portion of the gene). In many species, gene body methylation appears to be have both repressive and enhancer role. Recent methylome studies across phyla found that gene body methylation is mostly enriched for genes with moderate expression. In other words, genes that are either highly expressed or lowly expressed are depleted of gene body methylation. However, mammals do not seem to share this trait. Both human and mouse gene expressions do not tightly correlated with CG methylation in the gene body in protein coding genes. Furthermore, there is still no conclusive evidence that indicate gene body methylation plays a role in regulating gene expression.

On the other hand, gene body methylation for repetitive elements (such as retrotransposons) seems to be widely conserved across a diverse array of species. In many cases, heavy methylation across the repeat gene body results in stable silencing. Indeed, experiments that artificially remove DNA methylation within an organism results in dramatic
induction of repeat elements, and may lead to cell death \(^{51,52}\). It is thought that DNA methylation has evolved as a defense mechanism to silence foreign DNA such as from viruses, which are capable of invading a host cell for viral replication \(^{49}\). So although foreign viral DNAs have successfully integrated into host genomes over time, the cell has used DNA methylation as a way to provide genomic stability. On the same token, it has also been proposed that repeat elements are an important evolutionary driving force \(^{53,54}\). Retrotransposable elements are capable of “jumping” from its original locus and re-integrating in a random position of the genome and retrotransposon that randomly integrate into a gene’s regulatory element will directly affect its expression and may change the fitness of that cell. Indeed, mouse cells that are devoid of DNA methylation show increased expression of certain repeat elements, some of which can be found within transcripts (e.g. chimeric transcripts), reinforcing the idea for DNA methylation role in silencing repetitive elements to promote genomic stability \(^{55}\). Overall, DNA methylation within gene body of repetitive sequences seems to be functionally important and conserved across phyla.

**Non-CG methylation**

For many years, it was thought that CG methylation was predominant motif in mammalian cells. Indeed, this appears to be the case in most somatic cell types, but not in select cell types. For example, non-CG methylation accounts for about 25\% of all methylation in human embryonic stem cells (hESCs) \(^{47,56}\). However, not much is known about the functional relevance of non-CG methylation. Non-CG methylation is non-randomly distributed across the genome, and appears to be particular enriched in gene bodies \(^{47,56}\). In fact, non-CG methylation in the gene body appears to better correlate with gene expression. In hESCs, non-CG methylation has been suggested to enhance gene expression since the highly expressed genes tend to have more non-
Evidence suggests that non-CG methylation is primarily catalyzed by the de novo class of DNA methylation enzymes such as Dnmt3a and Dnmt3b. However, no study has teased apart the exact contributions from either enzyme. In summary, although non-CG methylation contributes approximately 25% of the global DNA methylation, its relevance to gene regulation still remains much a mystery.

5-hydroxymethyl-cytosine

Recent studies have identified a stable derivative of DNA methylation that is beginning to show emerging evidence for regulatory capacity. 5-hydroxymethylation (5hmC) is generated by oxidation of the 5'-methylcytosine (5mC) by a family of dioxygenases ten eleven translocation enzymes (Tets). Although the existence of 5hmC have been known for decades, it has only been recently brought into spotlight when it was found to be enriched in select cell types, such as Perkinje neurons and embryonic stem cells. Since then, there has been considerable interest in elucidating the potential roles for 5hmC and Tet enzymes. Interestingly, recent knockout studies in mouse demonstrate Tet1-deficient mice are viable, indicating Tet1 is dispensable for embryogenesis. Furthermore, Tet1 and Tet2 double knockout mouse are also compatible with development and post-natal defects appear to be variable. It has been postulated that Tet enzymes are involved in the demethylation pathway by labeling 5mC for downstream base-excision repair which eventually swaps out modified 5mC to unmodified cytosine or through passive demethylation by excluding Dnmt1 recognition for hemi-methylated DNA. Support for the role of 5hmC in demethylation comes from genome-wide Tet1 localization studies that find Tet1 is localized to regions of hypomethylated regions of high CpG density and Tet1-deficiency leads to increased global levels of DNA methylation.
Furthermore, new evidence is surfacing that implicate successive oxidation of the 5hmC to 5-formylcytosine and 5-carboxylcytosine\textsuperscript{71}. This principle has been leveraged to generate single-base resolution genome-wide hydroxymethylcytosine maps\textsuperscript{72,73}. Finally, it has been postulated that the last component of the 5hmC demethylation pathway involve TDG, the major base-excision repair enzyme. TDG knockout mouse is embryonic lethal and show dramatic increase of global 5mC, 5hmC, and 5fC levels in embryonic stem cells\textsuperscript{74-76}.

**Histone Modifications**

In the nucleus, genomic DNA is frequently found wrapped around proteins that help to fold and compact DNA. These core proteins, or histones, contain long N-terminal tails that are labile to extensive posttranslational modification\textsuperscript{77,78}. Some of the best characterized modifications involve methylation or acetylation of either lysine and arginine residues. Because these two residues carry a partially positive charge in aqueous environments, the negative charge of the acetyl-group is thought to neutralize the negatively charged phosphate backbone of DNA. This effectively reduces interactions between histone and DNA and is thought to reduces compaction and mediate transcription by allowing an open environment for recruiting RNA polymerase and other transcriptional machinery. Currently, over 50 histone modifications are known, but only a handful have been well characterized\textsuperscript{78}. Early global histone occupancy maps were huge undertakings, but have provided essential insight into the DNA chromatin environment and gene regulation\textsuperscript{79-83}. However, some histone modifications are rare and only localized in highly select regions of the genome, making it difficult to assess their occupancy and their function.

**Bivalency domains**
Intriguingly, within a single histone tail, multiple modifications may be found, sometimes with contradictory roles. One of the best examples of this phenomenon is co-localization of histone H3 lysine 4 trimethylation (a mark of active genes) and histone H3 lysine 27 trimethylation (a mark of repressed genes). This feature was first identified in embryonic stem cells, and since then has been these so-called bivalent domains have been proposed to be “poised” for quick activation or repression upon differentiation. Indeed, many bivalent genes in embryonic stem cells resolve to monovalency after differentiation. More recently, bivalent domains have been identified in other progenitor stem cells, suggesting perhaps bivalency is unique to stem cells for fine tuning the differentiation process.

Cross-talk with DNA methylation

In addition to co-occupancies of multiple histone modifications, other epigenetic marks are also often found localized together, indicating a clear cross-talk between multiple epigenetic mechanisms. For example, histone H3 lysine 36 trimethylation is often associated with high levels of gene body methylation, whereas H3K4me3 and H3K27me3 are often negatively correlated with DNA methylation. In neural stem cells, the mechanism for the latter process has been postulated to be carried out of Dnmt3a, whose DNA methylation activity inhibits recruitment of PRC2 complexes that catalyze trimethylation of H3K27. Thus, both DNA methylation and histones coordinate with each other to regulate gene activity.

Non-coding RNAs

From a historical perspective, non-coding RNAs did not enter center stage in light of the proteomics-centric emphasis on gene regulation, including roles of transcription factors, kinases,
and other regulatory enzymes. However, in recent years there has been dramatically increased interest in studying non-coding RNA. In part, this is due to advances in technology that allow facile assaying for non-coding RNAs, as well as increasing evidence that non-coding RNAs play important roles in gene regulation. Non-coding RNAs are a class of RNAs that do not encode proteins, but appear to have some regulatory function. Conceptually, non-coding RNAs work by complementary binding to signal increase or decrease downstream expression. Two broad classes appear to exist, one is long intergenic non-coding RNAs (lincRNAs) and another is the small-RNAs (microRNAs and piRNAs).

**MicroRNAs**

MicroRNAs (miRNAs) are short 18-22nt RNAs that appear to be primarily a repressive mechanism. It has been postulated that miRNAs can act to either by promoting mRNA degradation or by inhibiting translation. More recently, it has been shown that miRNAs appear to primarily act at the post-transcriptional level either through complementary binding or intronic or 3’UTR regions to mediate transcript degradation. MiRNAs are initially transcribed a longer pre-miRNA RNA, which is later processed by Dicer and Drosha enzymes to produce the mature functional miRNAs. Dicer and Drosha are essential for miRNA biogenesis and loss of either of these enzymes have severe repercussions on the cells. However, it is debated how loss of these enzymes affects cell survival and cell differentiation. On one hand, some have argued that deregulated miRNA network have caused the phenotypes, whereas others have argued that accumulation of pre- and pri-miRNAs lead to cellular toxicity (Fineberg et al., 2009). Regardless, loss of miRNAs directly affects global gene expression, suggesting miRNAs have substantial contribution in shaping the transcriptomes. Because miRNAs are short sequences, each miRNA
can complement many similar sequences across the genome and therefore can potentially regulate thousands of genes. It is becoming increasingly apparent that each cell type has a unique repertoire of signature miRNAs (Rossi et al., 2011; Marson et al., 2008), and these miRNAs contribute to cellular identity.

**piRNAs**

More recently, a class of Piwi-interacting RNA (piRNAs) have also been shown to be important for regulating gene expression. In mammals, piRNAs are most abundant in germ cells, but recently have also been identified to be embryonic stem cells and in neural cells. Within the genome, piRNAs are found in large clusters within intergenic heterochromatin regions and are thought to be transcribed in one long precursor form, then processed to smaller 28-30nt RNAs. Unlike miRNAs, the exact mechanism for how piRNAs are generated are still unclear, but the favored mechanism entails a “ping-pong” action, suggesting piRNA biogenesis utilizes a forward feedback-loop. In addition, piRNAs appear to be transcribed in a strand-specific manner, possibly indicating a necessary complementary action for silencing repeat sequences also found within heterochromatin.

**LincRNAs**

LincRNAs, like messenger RNAs, are transcribed in a precursor form which also undergoes processing such as splicing and tail poly-adenylation. LincRNAs range in size between 1kb and 10kb. Not much is known about lincRNAs except that some are well-conserved across species and are appear to contain regulatory promoters that follow conventional rules of transcription start sites. In fact, a wide repertoire of lincRNAs were originally predicted through
examine chromatin marks that did not correspond to any annotated protein coding regions but showed corresponding RNA transcripts. Much of our understanding of long non-coding RNA comes from studying the X-inactivation center, a cluster of non-coding RNAs that have important regulatory roles for X-inactivation. From these studies, we have learned a lot about the regulatory capacity of non-coding RNAs, such as base-complementary to facilitate X-chromosome pairing and inhibition of neighboring transcription. More recently, a comprehensive study that systematically knockdown individual lincRNAs in mouse embryonic stem cells show a majority of lincRNAs are key regulatory components of the pluripotency network. Together, these results demonstrate lincRNAs have important regulatory capacity on par with transcription factors also implicated in the pluripotency program.

**High-throughput Technologies Paving the Way for Epigenomics**

**Microarrays**

The advent of microarray technology provided a phenomenal method to measure multiple events in a single experiment. Expanding on classical complementary hybridization-based detection methods, the microarray platform relies on hybridization of fluorescently labeled DNA to pre-defined probes that uniquely represents portions of the genome. DNA probes are usually evenly spaced and attached to a solid surface, commonly referred to as a chip. Because of the limited size of the chip, only a finite number of probes can be placed onto a single chip. For experimental designs that attempt to exhaustively represent the genome by having probes found every few kilobases across the genome (so-called tiling arrays), multiple chips are required. Other experimental designs that focus on promoters alone may require less number of chips to fully represent all mammalian promoters. Microarray had clear advantages compared to previous
approaches in particular the ability to sample large portions of the genome in cost-effective and less time consuming manner.

**Next-generation sequencing**

In more recent years, high-throughput DNA sequencing has supplanted most microarray technologies for many reasons including improved high throughput, sensitivity, and accuracy. However, it is worth to mention that many laboratories are continuing to use microarray-based platforms primarily because of matured analytical tools and the costs are still less expensive for studies geared more for large sample sizes in the hundreds and thousands.

Sequencing offers many advantages over microarray platform including base-pair resolution and unbiased surveying of the genome. In general, all library construction protocols share fundamental commonalities. Since the end-goal is to generate short reads, the majority of library construction protocols share common procedures such as DNA or RNA fragmentation to a desired size distribution, adapter ligation, and PCR amplification. The PCR step is necessary because most library construction methods yield small amounts of DNA which may not be easily detected on the sequencer. On the other hand, the PCR step also remains one of the banes of library construction because PCR amplification introduces a variety of biases that confounds quantitative analyses. Indeed, several groups are working on methods for circumventing the PCR step in library construction, which will simplify library construction and data analysis in the future. So far, these procedures appear to only work with libraries with exorbitant starting materials, but low input libraries still have to undergo a non-trivial amount of PCR amplification. However, technology grows at a rapid pace and techniques are constantly improved upon.
In summary, the study of pluripotent stem cells has seen a dramatic increase in interest and improved knowledge in recent years. While early studies have been paramount to our modern understanding of the core pluripotency network, current technologies have taken stem cell research to an entirely new level of full genome-scale comprehensive analysis. Armed with a complete understanding of the nuances of stem cells, we will be able to devise better clinical approaches for stem cells.
References


Chapter 2
A panel of CG methylation sites distinguishes human embryonic stem cells and induced pluripotent stem cells

Kevin Huang1*, Yin Shen1*, Zhigang Xue1,3, Marina Bibikova2, Craig April2, Zhenshan Liu1,3, Linzhao Cheng6, Andras Nagy7, Matteo Pellegrini4, Jian-Bing Fan2, Guoping Fan1#

1Department of Human Genetics, Broad Stem Cell Research Center, David Geffen School of Medicine, University of California Los Angeles, 695 Charles Young Drive, Los Angeles CA, 90095 USA; 2Illumina, Inc. 9885 Towne Centre Drive, San Diego, CA 92121, USA; 3Translational Stem Cell Center, Tongji Hospital and Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China; 4Department of Molecular, Cell and Developmental Biology and 5Department of Pathology, University of California Los Angeles, California 90095 USA; 6Stem Cell Program in Institute for Cell Engineering and Division of Hematology, Johns Hopkins University, Baltimore, MD 21205, USA. 7Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 25 Orde Street, 5-1015-3 Toronto, Ontario, Canada M5T 3H7

*These authors made equal contributions to this study.
Abstract

Whether human induced pluripotent stem cells (iPSCs) are epigenetically identical to embryonic stem cells (hESCs) has been debated in the stem cell field. While some studies have shown iPSCs contain a unique DNA methylation signature, others have not found any CpG site consistently differentially methylated between iPSCs and hESCs, arguing instead that some sites tend to be more variably methylated than others sites. In this study, we analyzed DNA methylation patterns in a large number of human induced pluripotent stem cells (iPSCs, n=114) and embryonic stem cells (hESCs, n=155) and identify a panel of 82 CpG methylation sites that can distinguish iPSCs from hESCs with high accuracy. We showed that 12 out of the 82 CpG sites were subject to hypermethylation in part by DNMT3B. Moreover, we found that DNMT3B also directly contributes to silencing signature genes such as ZNF248 and TSPYL5 in iPSCs. Overall, we conclude DNMT3B is involved in a wave of de novo methylation during reprogramming, a portion of which contributes to the unique iPSC methylation signature. Moreover, these 82 CpG methylation sites would be useful to serve as biomarkers for distinguishing iPSCs and hESCs.
**Introduction**

DNA cytosine methylation is a major epigenetic factor that contributes to regulating important biological processes such as genomic imprinting, X-inactivation, and gene regulation\(^1\). DNA methylation is established and maintained by a family of DNA methyltransferases (DNMTs) including DNMT1, DNMT3A and DNMT3B\(^2\); deficiency in any Dnmt enzyme leads to embryonic death in mice. Additionally, aberrant DNA methylation are associated with human diseases such as cancer, immunodeficiency, and neurological disorders\(^3\). Collectively, these results demonstrate critical roles for DNA methylation in mammalian development.

DNA methylation is also important for both stem cell differentiation and cellular reprogramming. Although DNA methylation is dispensable self-renewal in ESCs, it is essential for differentiation\(^4,5\). Furthermore, dynamic DNA methylation changes during the process of ESCs differentiation are critical for lineage specification as a wave of *de novo* methylation takes place to silence pluripotency genes and establish tissue-specific methylation patterns\(^6-12\). During reprogramming, DNA methylation contributes to an epigenetic barrier. Indeed, selective promoter demethylation of pluripotent genes such as Oct4 and Nanog are associated with successful reprogramming of somatic cells to iPSCs\(^13-17\). In addition, inhibiting DNMTs activities with 5-aza-cytidine (AzaC), or knocking down DNMT1 promotes partially reprogrammed cells into fully reprogrammed state\(^13,14,18\). Meanwhile, a wave of *de novo* methylation also occurs during reprogramming whereby tissue-specific genes and partially methylated domains (PMDs) become hypermethylated\(^11,19,20\).

Human iPSCs have the characteristics of hESCs and many studies have investigated the similarities between human iPSCs and hESCs including genome stabilities\(^21-23\), transcriptome\(^24-26\), histone modifications\(^25\), and DNA methylation\(^19,20,27-31\). These studies revealed shared as well as
different properties between human iPSCs and hESCs. Specifically for DNA methylation, it has been reported that iPSCs acquire irregular methylation patterns while retaining some memory of somatic cells, thus exhibiting a methylation profile unique to iPSCs\textsuperscript{19,20,27-31}. However, because all previous studies differed in quantitation technique, genome coverage, and sample size, there has been a debate whether iPSCs possess a methylation signature that can be used to distinguish iPSCs from hESCs.

To address this question, we systematically compared methylation profiles in a large number of human iPSCs and hESCs from multiple labs to re-visit the question of whether iPSCs have a unique CpG methylation signature. We identified a panel of 82 CpGs that can distinguish iPSCs from hESCs with high accuracy. Remarkably, 12 of these signature CpG sites tend to be hypermethylated compared to both somatic cells and hESCs and localized in regions of low-CG density, suggesting these methylation signatures are not a form of residual somatic epigenetic memory. In addition, hypermethylation of these 12 sites are partially disrupted in DNMT3B-deficient iPSCs, consistent with the function of DNMT3B for \textit{de novo} methylation during reprogramming. Together, our results demonstrate a robust iPSC molecular signature that is partially a consequence of DNMT3B-mediated \textit{de novo} methylation during reprogramming.

\textbf{Results}

\textbf{A unique DNA methylation signature distinguishes human iPSCs from hESCs and somatic cells}

Previous reports have debated whether iPSCs exhibit a unique CpG methylation profile due to either residual somatic cell memory or aberrant methylation in select domains when compared to
hESCs. In this study, we investigated this issue by analyzing CpG methylation in a large number of pluripotent cell lines (n=269) from multiple labs, thus increasing the power of the statistical analyses. Using the Illumina Infinium beadchip assays, we first examined methylation profiles in 25 cell lines including 5 hESCs, 5 parental somatic cells, and 15 lines of iPSCs covering both iPSCs generated from vector-containing and vector-free methods. Hierarchical clustering analysis on all CpG sites measured by the beadchip assay demonstrated iPSCs are highly similar to hESCs, but distinctively different from somatic cells (Supplemental Fig. S1). To identify differential methylation between iPSCs and hESCs, we used a statistically stringent cutoff from Illumina’s custom model (see Methods) and required an absolute methylation difference (delta-beta) of 0.3. We found that the methylation profiles from 82 CpG sites in 66 genes can effectively group iPSCs separately from either hESCs or parental somatic cells (Fig. 2-1a, Supplemental Table S1). Gene ontology analysis revealed the signature genes were associated with epidermal cell differentiation and keratinization (Fig. 2-1b), suggesting that genes involved in tissue and cell differentiation contribute to the unique methylation pattern in iPSCs. Interestingly, the iPSC signature methylation pattern often resembled somatic cells, except at 12 CpG sites that appeared to be uniquely hypermethylated compared to somatic cells (Fig. 2-1a). This result suggested the iPSC methylation signature consists of both residual somatic memory and specific CpG sites subject to de novo methylation. Nevertheless, since pluripotent and somatic cells have well characterized differential methylation at pluripotency gene promoters such as OCT4 and NANOG, our study focused instead on examining alterations between iPSCs and hESCs.

Ultimately, a robust signature should be able to accurately discriminate between iPSC and hESC in independent datasets regardless of laboratory origin or quantitation method.
Although hierarchal clustering is one way to visualize how multiple samples are grouped, we turned to more robust and quantitative classification methods. To provide an unbiased estimate of predictive accuracy for cell type, we used a leave-one-out analysis where the support vector machine (SVM) learning model was fit on all but one sample and its prediction was related to the truly observed cell-type of the left-out sample (see Methods). Using other DNA methylation data available in the public domain, we consistently found our signature CpGs could significantly improve accuracy over randomly selected CGs for identifying iPSCs and hESCs. In general, we observed correct classification of greater than 95% of the samples. The results of the external validations are summarized in Table 2-1.

We found our panel of CG signature sites was able to accurately discriminate iPSCs and hESCs despite different studies having varying sample sizes. For example, we analyzed an independently derived Infinium 27k beadchip dataset that profiled DNA methylation in 20 iPSCs and 11 hESCs\textsuperscript{32}, and showed a classification accuracy of 96%. Another study using the same platform examined 42 and 115 normal iPSCs and hESCs\textsuperscript{33}. In this larger dataset, we could still accurately distinguish about 97% of the iPSCs and hESCs. Together, these results indicate that the methylation signature is robust to sample size sampling error.

Remarkably, our DNA methylation signature was also able to accurately discriminate iPSCs and ESCs regardless of quantitation platform or technique. A dataset employing the Infinium 450k beadchip system examined DNA methylation in 29 iPSCs and 18 ESCs. Interestingly, although only 70 of the 82 signature probes were shared between the two Illumina beadchip platforms, these 70 probes were still able to discriminate iPSCs from hESCs with 97% accuracy. Next, we tried to cross-reference our signature CpGs with publically available genome-wide bisulfite sequencing datasets. Although the reduced representation bisulfite
sequencing (RRBS) method is a cost-effective method to sample approximately \(\sim 3\) million CpGs in the human genome\(^{29,34}\), we found RRBS coverage had low overlap with the 82 signature CG sites (approximately 35 \([\sim 40\%]\) loci were detected) and not ideal for cross-referencing. We therefore turned to whole genome shotgun bisulfite sequencing datasets and curated a total of 5 iPSCs and 3 ESCs generated from three separate labs\(^{19,35,36}\). Strikingly, methylation quantitation of the 82 signature sites through bisulfite sequencing could also separate iPSCs from hESCs with 95\% accuracy. Taken together, we analyzed a total of 114 iPSCs and 155 hESCs collected from multiple labs, and these results indicate the identified CpG signature in our study is robust and can be broadly used to make a distinction between iPSCs and hESCs.

**Pair-wise comparisons of promoter CpG methylation between iPSCs and somatic cells reveal a wave of de novo methylation during reprogramming**

Because different somatic cells could have different tissue- and cell-specific methylation patterns\(^{12,20,37}\), we were interested in dissecting the precise methylation changes during the derivation of each iPS cell line. We therefore performed a pair-wise comparison in each CpG site between three pairs of iPSCs and their parental somatic cells generated in our lab. We first identified genes that exhibited statistically significant changes in methylation pattern using a delta-beta value \(\geq 0.3\) or \(\leq -0.3\) cutoff. Our data indicated that approximately 7-14\% of gene promoters underwent methylation changes during direct reprogramming (Fig. 2-2a). Surprisingly, we found that the number of gene promoters exhibiting an increase in methylation was 3.5-6 folds more than the number of gene promoters showing a decrease in methylation. Consistently, high performance liquid chromatography mass spectrometry (HPLC-MS) showed
an overall increased level of DNA methylation in iPSCs when compared to parental somatic cells (Fig. 2-2b).

Furthermore, by cross-referencing gene expression profiles between iPSCs and parental somatic cells, we found that approximately 60-75% of genes that were subject to de novo methylation showed a significant reduction of gene expression or no expression in iPSCs (Fig. 2c and Supplemental Table S2). The increased methylation levels at promoter CpG sites in iPSCs were confirmed by conventional bisulfite sequencing analysis (Supplemental Fig. S2). In total, 151 genes showed hypermethylation at promoter CpG and also suppressed in all three pairwise comparisons. Gene ontology analysis indicated that these silenced genes were enriched for the genes required for specific functions such as immune system process and receptor activity (Fig. 2-2d). Consistently, these silenced genes were depleted from genes involved in housekeeping functions such as intracellular membrane organelle, cellular metabolic process, and regulation of transcription (Fig. 2-2d). Our analysis suggested that de novo methylation activities contribute to the silencing of genes involved in specialized cellular function and differentiation pathways during the conversion of somatic cells into human iPSCs.

In addition to global hypermethylation, a portion of the methylation signature was also uniquely hypermethylated (n=12) in iPSCs but hypomethylated in both somatic cells and a portion of hESCs (Fig. 2-1a). Remarkably, these 12 sites were consistently hypermethylated in iPSCs found in other datasets (Fig. 2-3a), confirming these sites tended to show unidirectional differential methylation. However, on several occasions we found heterogeneity in the methylation level at these 12 loci in hESCs (Supplemental Fig. S3). Interestingly, closer inspection of hESCs from the Nazor et al. dataset revealed cell lines from the CM, ESI, FES, SIVF, UC06, and MIZ series tended to be hypermethylated whereas the HES, WA, MEL, and
MIV series tended to be hypomethylated (Supplemental Fig. S3c). These differences did not appear to be laboratory dependent since cell lines such as WA09 (also referred to as H9) cultured in four separate labs were consistently hypomethylated at these sites (Supplemental Fig. S4).

**Hypermethylation by DNMT3B contributes to the panel of methylation signature**

Because Dnmt3B is more dramatically up-regulated in iPSCs when compared to levels of Dnmt1 and Dnmt3A, we hypothesized DNMT3B may play a major role for de novo methylation in human iPSCs. To test this hypothesis, we generated iPSCs from skin fibroblasts of Immunodeficiency, Centromere instability, and Facial anomalies (ICF) syndrome patients carrying double heterozygous point mutations in the catalytic domain of DNMT3B, and mapped the methylome for two ICF iPSC cell lines at base-pair resolution via whole genome shotgun bisulfite sequencing (Huang et al., in preparation). We achieved an average coverage of 10X and 4X per strand for ICF1#1-iPSCs and ICF1#2-iPSCs (Huang et al., in preparation). By mapping the 27K CpG sites arrayed on Illumina Beadsarrays or taking the whole methylome, we found that ICF-iPSCs consistently clustered away from normal hiPSCs and hESCs (Supplemental Fig. S5), indicating a dramatic change in the DNMT3B-deficient iPSC methylome.

By cross-referencing other whole genome bisulfite sequencing datasets, we confirmed our 12 hypermethylation signature sites were hypomethylated in parental somatic cells, but then hypermethylated in iPSCs (Fig. 2-3b). In ICF-iPSCs, methylation levels at these 12 sites were generally reduced (Fig. 2-3b), suggesting a role for DNMT3B in de novo methylation at these 12 targets. Remarkably, all 12 CpG sites were located in regions of low CG-density and mostly devoid of histone marks (Fig. 2-3c). We also examined whether these signature sites had any
similarities in genetic regulatory elements, but found no significant enrichments in TF binding motifs or direct TF binding. This result suggests promoters with low CG density are somehow more susceptible to de novo methylation by DNMT3B during cell reprogramming.

We next generated expression profiles via performed RNA-seq in ICF iPSCs and control pluripotent stem cells and asked whether any signature genes were altered in DNMT3B-deficient iPSCs. Of the total 136 and 213 up- and down-regulated genes (FDR<1% and fold change > 1.5) in ICF iPSCs, 4 signature genes overlapped including CSRP1, ZNF248, TSPYL5, and ASCL2. Remarkably, ZNF248 and TSPYL5 showed selective promoter hypomethylation concomitant with increased expression, suggesting these genes may be directly regulated by DNMT3B (Fig. 2-3d and Supplementary Fig. S6). Together, we conclude DNMT3B plays a role in targeting some signature genes.

Discussion

It was not clear whether human iPSCs have distinct transcriptomes and methylomes when compared to hESCs. Although one initial study reported the presence of iPSC-specific gene expression in a small number of iPSCs\textsuperscript{24}, several other studies argued that at least on the individual gene expression level, there are large variations between separate datasets\textsuperscript{25,26}. Recognizing the limitations for analyses based on individual genes, we previously utilized weighted gene co-expression network analysis (WGCNA) to identify functional modules that are distinct between iPSCs and ESCs\textsuperscript{38}. We further showed the expressions of these functional modules were inversely correlated with the level of DNA methylation in gene promoters, suggesting there were specific methylation changes in iPSCs. These findings prompted us to re-
investigate the contentious issue of whether human iPSCs have a unique methylation signature to
distinguish iPSCs from both hESCs and somatic cells.

By analyzing the CpG methylation in 269 cell lines of human pluripotent from many
different labs, we identified a unique methylation signature for iPSCs compared to hESCs and
somatic cells. Because iPSCs exhibit a significant increase in genome-wide methylation when
compared to parental somatic cells, we suspected that de novo methylation played an important
role in establishing a unique iPSC methylation signature. By comparing methylation patterns in
mutant ICF1-iPSCs, we indeed found altered methylation signatures, suggesting DNMT3B
contributes to de novo methylation during reprogramming. In particular, we identified 12
signature CpGs (out of the 82 CpG sites) that undergo DNMT3B-mediated de novo methylation.

Our methylation signature is different from what was previously identified by either
microarray or high-throughput sequencing analysis. Earlier studies suffered primarily from
limited sample sizes due to the costly approach to measure genome-wide DNA methylation
levels on a comprehensive scale. Several previous studies using RRBS attempted to verify
reported signatures in the literature and found a lack of reproducibility\textsuperscript{39,34}, arguing instead for
variations in iPSCs. As RRBS only covers \textasciitilde 10\% of the human CpG sites and biased for regions
of high CpG density, it is possible that the coverage could not fully detect the regions that were
consistently different in iPSCs. Indeed, a significant portion of our 82 signature resided in non-
CG islands. A more recent study by Ruiz et al. using the bisulfite sequencing padlock probe
(BSPP) system identified 9 signature genes that distinguish hESCs and hiPSCs\textsuperscript{39}. On average,
BSPP covers 500,000 thousand CpGs in the human genome\textsuperscript{40} (\textasciitilde 1\% of all human CpG sites); however, these sites have low overlap with the Infinium 27k array (\textasciitilde 25\% shared sites within
100bp). Interestingly, TCERG1L, the only identified signature gene that overlaps with the
Illumina 27k Beadarray, was also identified as a signature gene in our study. This indicates at least one gene can be validated in a separate dataset using a different platform. Moreover, when we compared our list of signature CG sites with other signatures in the literature, we found minimal overlap\(^{19,20}\). However, it is still unclear whether this low overlap is due to incompatible coverage or lack of sample size for robust delineation of an accurate signature. For example, Lister et al. (2011) initially identified hundreds of CG-DMRs in iPSCs, of which only a small fraction could be confirmed in multiple cell lines, suggesting the number of sites is gradually reduced as sample size becomes larger. By contrast, although our study identified a methylation signature using 25 cell lines, we were able to validate these signatures in 249 other samples, demonstrating our signature comprises a core set of CpG sites that can reliably distinguish iPSCs, hESCs, and somatic cells. Overall, we suggest that although a definitive signature whereby a given site is always differentially methylated between the two cell types may not exist, a panel of CpG sites representing loci that tend to be differentially methylated is sufficient to segregate iPSCs and hESCs. The identification of a panel of CpG methylation signatures in iPSCs will be useful to serve as a molecular biomarker for classifying iPSCs in the future.

**Methods**

**Derivation and cultures of human iPSCs and hESCs**

The production of human iPSCs follows the protocol described by Takahashi *et al.* (2007) and Yu *et al.* (2007) using retroviruses expression OCT4, SOX2, KLF4, and c-MYC or OCT4, NANOG, KLF4 and LIN-28. hESC cells were maintained in DME supplemented with 20% KSR, nonessential amino acids (Invitrogen), L-Glutamine (Mediatech), Penn/Strep, 2-mercaptoethanol with a feeder layer of MEFs as previously described (Shen *et al.*, 2006). For
both gene expression and methylation analysis studies, the hESCs were passaged onto feeder free gelatin coated plates twice before harvesting RNA and DNA. RNA was isolated using Trizol (Invitrogen) while DNA was isolated using PureLink™ genomic DNA purification kit (Invitrogen).

ICF1-1 and ICF1-2 iPSCs were derived from two different lines of ICF patient fibroblasts carrying non-conserved double heterozygous mutations in the catalytic domain. ICF1 iPSCs can form teratoma and differentiate into three germ-layer lineages, demonstrating the pluripotency of these iPSCs. The details of ICF iPSC characterization will be reported in a separate manuscript by Huang et al. (in preparation).

**DNA methylation profiling with Illumina Infinium assays**

We have used the HumanMethylation27 DNA Analysis BeadChip from Illumina, Inc. (San Diego, CA) to interrogate 26,837 highly informative CpG sites over 14,152 genes. The human DNA sequence was based on the NCBI CCDS database (Genome Build 36) as described by the manufacturer. The experimental procedures of bisulfite conversion of genomic DNAs, hybridization of HumanMethylation27 BeadChips, and extraction of raw hybridization signals follow manufacturer’s instruction. Data analysis was performed with the BeadStudio software from Illumina, Inc. The assays were done in technical duplicates for each cell line, and exhibited very high correlations (average $R^2 = 0.998$).

**Whole-genome gene expression analysis: humanht-12 beadchip and Agilent 44k**

Gene expression microarrays were performed with Illumina Whole-Genome expression microarrays (HumanRef-8 v3.0 Expression BeadChip, Illumina, Inc. San Diego, CA) or Agilent 44K Whole Human genome arrays (G4112A; Agilent Technologies) using the suggested
protocol. BeadStudio Software and Bioconductor package were used for data processing and analysis.

**Clustering analysis of methylation data**

The cluster analysis of methylation data was performed using Methylation Module v1.0 in BeadStudio (Illumina, Inc.) according to the manufacturer’s manual. Briefly, average signals of built-in negative control were used as the background value to normalize the methylation signals. Outliers are removed by using the median absolute deviation method. Methylation levels of individual loci in individual samples and sample groups were presented as beta values, which are estimated by calculating the ratio of intensities between methylated and unmethylated alleles. Differential methylation analysis algorithms and error models inherent in BeadStudio were used following the Illumina Custom Model. Briefly, the model assumes beta values are normally distributed, and p-values are defined as difference in mean and variance between beta values from two conditions (z-score). In general, differential methylation (or delta-beta) of >|0.3| are considered highly significant. Samples with differences of delta-beta value >|0.3| were therefore selected and subjected to the clustering analysis using cluster methods built in the BeadStudio software (Illumina, San Diego, CA).

**Signature Validation and Data analysis**

We performed quantile normalization for the Illumina 27k Infinium dataset from Chou et al. (2011) and Nazor et al. (2012). For the Illumina 450k Infinium dataset, we first performed peak correcting using the IMA package for R\(^41\), followed by quantile normalization.

For whole genome shotgun bisulfite data, we firstly mapped all raw reads to the hg18 genome using BS-Seeker\(^42\). Next, we identified all CGs that fell within 5bp of each signature CpG site. If no CGs were found, we searched for CGs that were at most 100bp flanking the signature CpG
site. For cell lines with no CGs found within this vicinity (i.e. missing data), we imputed the data point using nearest neighbor method only if a single data point was missing for a given gene. We did not perform any normalization on the bisulfite sequencing data, since these signals are inherently highly quantitative. For assessing promoter methylation (Fig. 3d), we took the average methylation in the 500bp flanking the TSS.

**Support vector machine**

Support vector machines are a type of classifier method based on hyperplanes in a high dimensional space in which samples can be separated by their distance from the planes. The R package ‘e1071’ was used to train the datasets and predict accuracy of classifying iPSCs or ESCs (linear kernel function). We used two-thirds data as training set, and the rest (33%) as test data. Leave-one-out cross-validation was performed using the ‘cross’ function (the parameter was set to equal the total number of samples) inherent in the R package. Accuracy was calculated as the portion of correctly classified samples in the SVM model. To generate the background (or null) distribution of a random signature, we calculated the accuracy of re-iteratively sampled 82 random sites (n= 20,000) and reported the mean of the distribution. False discovery rate (FDR) was determined by finding the portion of the distribution that was greater than the observed accuracy of the signature sites.

**Bisulfite Conversion and Sequencing**

Bisulfite conversion was performed as described⁶. Briefly, we digested genomic DNA with BglII overnight. Digested DNAs were then incubated with a sodium bisulfite solution for 16 hours. Bisulfite treated DNA was then desalted and precipitated. We used 1/10 of precipitated DNA for each PCR. For PCR, we used nested primers to generate our products. PCR products were gel purified and used for Topo Cloning (Invitrogen).
Gene ontology analysis

Gene Ontology analysis was performed using DAVID Bioinformatics Resource (http://david.abcc.ncifcrf.gov/).

Accession Number:

Microarray data has been deposited at GEO (Gene Expression Omnibus) under accession number GSE42043.

Acknowledgements

We thank Rudolf Jaenisch for generously providing DNA samples of iPSCs for this study. Funding for this work is from NIH R21 NS072924 and California Institute for Regenerative Medicine (CIRM RC1-0111-01). Kevin Huang is supported by the California Institute of Regenerative Medicine Training Grant (TG2-01169).

Figure legends

Figure 2-1. A unique signature distinguishes iPSCs, hESCs, and somatic cells.

A. Heatmap representation of the 82 signature CpGs in 15 iPSCs, 5 hESCs, and 5 somatic cells. The blue box highlights CpGs that underwent de novo methylation in iPSCs. B. Gene ontology of the signature genes (n=66).

Figure 2-2. Pair-wise comparison between parental somatic cells and iPSCs reveals alterations of promoter methylation in reprogramming and correlation with gene expression
A. Global view of DNA methylation changes during the reprogramming of parental somatic cell lines to iPSCs. Using the delta-beta $>0.3$ (increase in methylation) or $<-0.3$ (decrease in methylation), we compiled the pie chart after comparison of 26,837 CpG sites in 14,512 genes for each pair of somatic cells and iPSCs. B. Global methylcytosine levels as measured by HPLC-MS. C. Cross-reference analysis of genes showing increased methylation in iPSCs with their status of gene expression changes between iPSCs and somatic cells. D. Gene ontology analysis of genes with de novo methylation and decreased expression. The GO term is on the y axis, and the p value indicating significance is on the x axis. P-values of GO terms which are overrepresented in the dataset are colored in red; p-values of underrepresented or depleted GO terms are colored in green ($p < 0.05$).

Figure 2-3. A wave of de novo methylation during reprogramming by DNMT3B.

A. Boxplot of the average methylation levels for the 12 sites that tend to be hypermethylated in iPSCs. ‘+’ sign denotes outliers. ESC$_1$ denote hESCs from the CM, ESI, FES, SIVF, UC06, and MIZ series, and ESC$_2$ denote hESCs from HES, WA, MEL, and MIV series. B. Heatmap of the 12 CpGs through bisulfite sequencing. Legend represents raw methylation levels. C. Bar graph of the percentage of sites positive for the labeled attributes. D. Genome browser view of expression and DNA methylation levels at the signature gene, ZNF248. Note selective hypomethylation at the proximal promoter associated with increased gene expression.
Figure 2-1. A unique signature distinguishes iPSCs, hESCs, and somatic cells.
Figure 2-2 Pair-wise comparison between parental somatic cells and iPSCs reveals alterations of promoter methylation in reprogramming and correlation with gene expression.
Figure 2-3. A wave of de novo methylation during reprogramming by DNMT3B.
Table 2-1. Classification accuracy of signature genes by SVM

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Platform</th>
<th>No. samples (iPS/ES)</th>
<th>Accuracy random 82 (%)</th>
<th>Signature 82 (%)</th>
<th>FDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huang et al.</td>
<td>Illumina 27k</td>
<td>20 (15/5)</td>
<td>81</td>
<td>100</td>
<td>4.55</td>
</tr>
<tr>
<td>Cheng et al.</td>
<td>Illumina 27k</td>
<td>31 (20/11)</td>
<td>69</td>
<td>96</td>
<td>0.01</td>
</tr>
<tr>
<td>Nazor et al.</td>
<td>Illumina 27k</td>
<td>163 (45/118)</td>
<td>83</td>
<td>97</td>
<td>0.01</td>
</tr>
<tr>
<td>Nazor et al.</td>
<td>Illumina 450k</td>
<td>47 (29/18)</td>
<td>82</td>
<td>97</td>
<td>0.31</td>
</tr>
<tr>
<td>Lister et al.; Laurent et al., Chodavarapu et al.</td>
<td>WGBS</td>
<td>8 (5/3)</td>
<td>68</td>
<td>94</td>
<td>7.10</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>269 (114/155)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FDR was determined by computing the accuracy of randomly selected 82 CG sites (n=20,000) to generate a background (or null) distribution, then finding the portion of the distribution that was greater than the observed accuracy of the signature sites.
References


Shi, Y. *et al.* A combined chemical and genetic approach for the generation of induced pluripotent stem cells. *Cell Stem Cell* **2**, 525-528, doi:S1934-5909(08)00225-7


Chapter 3
Functional Modules Distinguish Human Induced Pluripotent Stem Cells from Embryonic Stem Cells

Anyou Wang,1 Kevin Huang,1 Yin Shen,1 Zhigang Xue,2 Chaochao Cai,1,3 Steve Horvath,1,3 and Guoping Fan1,2

It has been debated whether human induced pluripotent stem cells (iPSCs) and embryonic stem cells (ESCs) express distinctive transcriptomes. By using the method of weighted gene co-expression network analysis, we showed here that iPSCs exhibit altered functional modules compared with ESCs. Notably, iPSCs and ESCs differentially express 17 modules that primarily function in transcription, metabolism, development, and immune response. These module activations (up- and downregulation) are highly conserved in a variety of iPSCs, and genes in each module are coherently co-expressed. Furthermore, activation levels of these modular genes can be used as quantitative variables to discriminate iPSCs and ESCs with high accuracy (96%). Thus, differential activations of these functional modules are the conserved features distinguishing iPSCs from ESCs. Strikingly, the overall activation level of these modules is inversely correlated with the DNA methylation level, suggesting that DNA methylation may be one mechanism regulating the module differences. Overall, we conclude that human iPSCs and ESCs exhibit distinct gene expression networks, which are likely associated with different epigenetic reprogramming events during the derivation of iPSCs and ESCs.

Introduction

Induced pluripotent stem cells (iPSCs) produced from somatic cells by overexpressing key transcription factors closely resemble embryonic stem cells (ESCs) in many aspects, including cell morphology, chromatin modifications, and differentiation potency [1-6]. Human iPSCs have become a powerful tool for biomedical research and may provide a promising alternative for cell-replacement therapies [7-9]. However, regardless of parental cell lineages or reprogramming techniques, several studies have shown that iPSCs are different from ESCs at the level of RNA transcription, leading to a debate regarding whether iPSCs are truly similar to ESCs [10-14].

It is suggested that transcriptome changes between human ESCs and iPSCs arise from different culture conditions or different laboratory practices [1-2,10-12]. This hypothesis is supported by cluster analysis of gene expression profiling from different research groups [11,12], in which iPSCs and ESCs derived from individual research labs tend to be clustered together into a lab-specific pattern [11,12]. However, these analyses simply merged gene expression data generated from different labs without removing batch effects, which may significantly mislead the conclusions derived from independently measured microarray data [15]. These lab-specific gene expression patterns between iPSCs and ESCs may need more thorough re-examination.

Several studies have attempted to identify individual iPSCs differentially expressed genes between iPSCs and ESCs. One study reported a total of 294 differentially expressed genes between human iPSCs and ESCs, suggesting that iPSCs have a unique expression signature [13]. However, these 294 individual gene signatures are not conserved in different iPSCs after independently re-examining the same database by several groups [11,12,14]. This suggests that unique and reliable gene expression signatures distinguishing iPSCs and ESCs still remain elusive.

In contrast to individual gene expression signatures that are less conserved in various iPSCs as discussed above, certain functional groups have been consistently found to be altered between ESCs and iPSCs [16,17]. For example, functional groups involved in development, transcription, immune response, and enzyme activities for metabolism have been frequently found in recent studies [16,17]. Functional groups (modules) are believed to be stable units in systems biology because the overall function of a module can remain the same, whereas individual gene expression can be changed or replaced by other genes with similar redundant functions. Potentially, functional modules can more effectively reveal consistent differences between iPSCs and ESCs than individual gene signatures.

1Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California.
2Stem Cell Research Center, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China.
3Department of Biostatistics, School of Public Health, University of California, Los Angeles, California.
Here, we utilized a systems biology method, weighted
gene co-expression network analysis (WGCNA), to analyze a
large set of genomewide gene expression profiles of typical
human iPSCs and ESCs. Our analysis revealed that iPSCs are
inherently different from ESCs at the module level. In par-
ticular, we identified 17 functional modules primarily func-
tioning in transcription, development, immune response,
and metabolism that distinguish iPSCs from ESCs. We fur-
ther demonstrated that differentially expressed functional
modules are associated with different DNA methylation
profiles between human iPSCs and ESCs.

Materials and Methods

Microarray data

Microarray data for iPSCs and ESCs were gathered from
previously published data deposited in GEO (www
.ncbi.nlm.nih.gov/geo/). The collected database includes data of various iPS cells such as those derived from different
sets of gene combinations and different cells, even different
species, human and mouse. Concerning the well-known
data variations derived from different microarray platform,
we focused on data generated by Affymetrix platform.
However, for validation, we also include one set of data
from Illumina microarray platform. The following data sets
were extracted, including human genome U133 Plus 2.0 ar-
ray, GSE12590, GSE14711, GSE15176 GSE15148, GSE1693,
GSE16545, and GSE39665; Affymetrix Mouse Genome Array,
GSE 14012, GSE18086, GSE10871, and GSE15267; and Illu-
mina, GSE16062. Three new available datasets were also
included: GSE27280, GSE24355, and GSE23583.

DNA methylation profiling with Illumina

Human Methylation DNA Analysis BeadChip from Illu-
mina, Inc. (San Diego, CA), was used to interrogate 26,837
highly informative CpG sites over 14,152 genes for 10 samples,
5 iPSCs (HNPC8IPS, HNPC9IPS, HNPC10IPS, HNPC11IPS,
and IMR90IPS), and 5 ESCs (HSF6, H1, H9, HSF1, and Hues7).
The experiment was performed following procedures based
on the manufacturer’s instructions, including bisulfite con-
version of genomic DNAs, hybridization, and extraction of
raw hybridization signals. BeadStudio software from Illumina,
Inc., was used to analyze the methylation data.

Gene expression data analysis

The microarray data were analyzed using R (www
.r-project.org/), the preliminary array quality assessment
with affyQCReport package, the background adjustment and
normalization with affy package, and the gene expression
values estimation with limma package. Because these mi-
croarray data were generated by different research groups,
the batch effect should be filtered out before combining these
microarray datasets. An algorithm called ComBat [15], which
runs in R environment and uses parametric and nonpara-
metric empirical Bayes frameworks to adjust microarray data
for batch effects, was used to adjust the final gene expression
values for all datasets.

After filtering the outlier chips by the preprocessing
function from our network software, WGCNA [18], we had a
total of 47 chips for network analysis: 34 iPSCs and 13 ESCs.
These 34 remaining iPSCs samples were generated by the
most stringent methods and their biological properties close
to human ESCs.

Network construction and module identification

The network was constructed by using WGCNA as we
previously described [18]. Briefly, WGCNA measure any
gene pair (i,j) similarity Si,j as defined below [19–22].

\[
S_{ij}^{(corr)} = \frac{1 + \text{cor}(z_i, z_j)}{2}
\]

where \( z_i \) and \( z_j \) are the gene expression of genes \( i \) and \( j \) across
multiple microarray samples.

The similarity was measured continuously with a power \( \beta \)
as a weight to obtain the weighted adjacency \( x_{ij} \) for any gene pair as

\[
x_{ij} = S_{ij}^{(\beta)}
\]

where \( \beta \) can be chosen using the scale-free topology criterion.
Since \( \log(\text{adj}) = \beta \times \log(\text{sim}) \), the overall network adjacency is
linearly correlated with the co-expression similarity on a
logarithmic scale. The adjacency matrix \( A = [ x_{ij} ] \) constructs a
weighted network.

The network modules are defined as cluster branches
derived from hierarchical clustering based on the network
proximity as input. The proximity is defined by the topo-
logical overlap measure [18,20–22] of connection strengths of
all possible gene pairs collected in the adjacency matrix \( A \)
described above.

The network and module membership

The network membership is measured by the network
eigengene based connectivity, \( K_i \) [23–26]

\[
K_i = \text{cor}(X_i, E)
\]

where \( X_i \) is the expression profile of the gene \( i \) and \( E \) is the
eigengene of the network as defined below.

\[
E = V_1
\]

where \( V_1 \) is the singular vector in \( V \) below corresponding to the
largest absolute singular value in \( D \) below

\[
X = UD(V)^T
\]

where \( X \) is the \( n \times m \) matrix of standardized expression
profiles of the \( n \) genes in the network/network-module
across \( m \) samples, \( U \) is an \( n \times m \) matrix with orthogonal
columns, \( D \) is an \( m \times m \) diagonal matrix of singular values,
and \( V \) is an \( n \times m \) orthogonal matrix of singular
vectors.

Key node identification

The connectivity considers both the network topology and
the eigengene-based connectivity as we previously reported
[26,27] and as defined below.
MODULiS IN iPSCs AND ESCs

\[
\text{score} = \frac{d_{i}}{d_{\text{max}}} + 2\times \text{cor}(x_{i}, E)
\]

where \(d_{i}\) represents the \(i^{\text{th}}\) node degree that measures the total connectivity of the \(i^{\text{th}}\) node, and \(d_{\text{max}}\) represents the maximum degree of a node in the network. \(\text{cor}(x_{i}, E)\) is the absolute value of Pearson correlation coefficient, where \(x_{i}\) is a vector of gene expression of \(i^{\text{th}}\) node, and \(E\) eigengene of the network. We put twice weight on eigengene-based connectivity because our network is highly connected in topology and our data showed almost equal importance on first and second components.

Support vector machines. Support vector machines (SVMs) [28] are a set of related machine learning methods for classifying datasets based on hyperplanes in a high or infinite dimensional space in which samples of a cluster can be separated with the largest distance to others. The R package e1071 was used to train our datasets and predict the accuracy of module-based separations in this study. We used 70% samples as training set, and the rest (30%) as test data. The accuracy was calculated by measuring both average percentage and kappa value (a value for measuring agreement) after randomly sampling 3,000 times for each module combination. The module combination starts with 1 module to 2 modules, 3 modules until 17 modules in 17 module set (Table 1, Fig. 3C), and begins from 1 module to 2 modules, until 4 modules in 4 super-module set (Fig. 3D).

**Table 1. Total 17 Modules Differently Expressed in Human Induced Pluripotent Stem Cells and Embryonic Stem Cells**

<table>
<thead>
<tr>
<th>Module no.</th>
<th>Module color</th>
<th>(P) value</th>
<th>Nodes</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Blue</td>
<td>2.50E-28</td>
<td>110</td>
<td>Gene expression and RNA metabolism</td>
</tr>
<tr>
<td>11</td>
<td>Lightyellow</td>
<td>4.12E-21</td>
<td>13</td>
<td>RNA binding/lysosomal lumen acidification</td>
</tr>
<tr>
<td>8</td>
<td>Grey</td>
<td>4.01E-20</td>
<td>30</td>
<td>ubiquitin-dependent protein catabolic process, glycine dehydrogenase (decarboxylating) activity</td>
</tr>
<tr>
<td>3</td>
<td>Brown</td>
<td>8.50E-20</td>
<td>99</td>
<td>transferase activity, signaling transduction</td>
</tr>
<tr>
<td>9</td>
<td>Lightcyan</td>
<td>1.67E-18</td>
<td>15</td>
<td>catalytic activity/nucleoside triphosphate adenylosuccinate lyase</td>
</tr>
<tr>
<td>10</td>
<td>Lightgreen</td>
<td>8.91E-16</td>
<td>14</td>
<td>peptide antigen-transporting ATPase activity</td>
</tr>
<tr>
<td>7</td>
<td>Greenshadow</td>
<td>3.74E-15</td>
<td>37</td>
<td>glutathione transferase activity, myeloid progenitor cell differentiation</td>
</tr>
<tr>
<td>16</td>
<td>Turquoise</td>
<td>1.82E-13</td>
<td>163</td>
<td>DNA repair, transcription</td>
</tr>
<tr>
<td>12</td>
<td>Midnightblue</td>
<td>2.95E-13</td>
<td>15</td>
<td>acetyltransferase activity, enoyl-[acyl-carrier-protein] reductase activity</td>
</tr>
<tr>
<td>6</td>
<td>Green</td>
<td>3.72E-13</td>
<td>40</td>
<td>GMP-activated cyclic-nucleotide phosphodiesterase activity, negative regulation of lymphocyte differentiation</td>
</tr>
<tr>
<td>4</td>
<td>Cyan</td>
<td>1.68E-24</td>
<td>32</td>
<td>RNA splicing, vitamin B6 metabolic process</td>
</tr>
<tr>
<td>15</td>
<td>Tan</td>
<td>2.88E-21</td>
<td>22</td>
<td>acute inflammatory response, membrane organization and biogenesis regulation of cell growth</td>
</tr>
<tr>
<td>13</td>
<td>Pink</td>
<td>4.68E-17</td>
<td>30</td>
<td>Cell surface/regulation of developmental process</td>
</tr>
<tr>
<td>17</td>
<td>Yellow</td>
<td>9.04E-17</td>
<td>41</td>
<td>mediator complex/regulation of adenosine receptor signaling pathway</td>
</tr>
<tr>
<td>1</td>
<td>Black</td>
<td>1.04E-14</td>
<td>57</td>
<td>glycoprotein-N-acetylactosamine-3-beta-galactosyltransferase activity</td>
</tr>
<tr>
<td>14</td>
<td>Salmon</td>
<td>1.19E-14</td>
<td>20</td>
<td>catalytic activity, single-stranded DNA specific endodeoxyribonuclease activity</td>
</tr>
<tr>
<td>5</td>
<td>Darkred</td>
<td>6.99E-13</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

**Results**

Human iPSCs and ESCs exhibit distinctive transcriptional profiles

Recent studies reported that distinctive iPSC expression profiles are lab-specific patterns [11,12]. However, analyses did not take into account of experiment effects that may confound the conclusions. To deter these profiling are intrinsic properties in various iPS random across different conditions, we analyzed all iPSC and ESC gene expression datasets deposited in GEO database (www.ncbi.nlm.nih.gov/geo/) [7] (Materials and Methods section). These datasets data of various iPSCs resources such as virus-integrated iPSCs, vector-free iPSCs, and protein-directed reprogramming iPSCs (Materials and Methods section). These datasets generated by stringent methods are well characterized and show similar biological properties to human ESCs. To compare all collected gene expression datasets dated from different labs, we first filtered potential based on low inter-array correlation, followed by normalization, and finally batch effect removal (M and Methods section). Therefore, the different expression patterns expressed between these iPSCs or should represent typical profiling variations between iPSCs and ESCs.
FIG. 1. Human iPSCs express distinctive transcription profiles compared with ESCs. Typical iPSCs and ESCs gene expression profiles were analyzed by both cluster analysis and between-group lab-specific patterns were gone after removing batch effects. Cluster analysis (A, B) of samples before (A) and after (B) removing batch effects. (C) Between-group analysis of same samples and before (left panel) and after (right panel) removing batch effects. Samples were labeled by GEO deposited number. iPSCs, induced pluripotent stem cells; ESCs, embryonic stem cells.

Without removing batch effects, cluster analysis of these expression data shows lab-specific groupings (Fig. 1A) as previously reported; however, the same analysis with removed batch effects revealed cell-type-specific profiling (Fig. 1B). In other words, ESCs are mostly separated from iPSCs regardless of lab origin; only 2 ESC samples were misgrouped (Fig. 1B). To determine whether the misgrouped samples were caused by computational limitations of cluster analysis, we applied between-group analysis, a high sensitivity multivariate analysis [32], to discriminate the samples. Consistently, between-group analysis of the same samples revealed 2 clearly separated groups after batch effect removal (Fig. 1C, right panel). This indicated that batch effects caused the lab-specific groupings, and that iPSCs indeed show distinct gene expression profiles compared with ESCs.

Gene networks are differentially expressed in iPSCs and ESCs

Because individual gene signatures distinguishing iPSCs and ESCs could not be extracted [11,12], we turned to gene network analysis to examine consistent functional module differences between iPSCs and ESCs by employing WGCNA [18] (see Materials and Methods section).

WGCNA analysis of the differentially expressed genes (Materials and Methods) produced a network significantly altered between iPSCs and ESCs (Fig. 2, P < 3.305E-08). This network contains 751 nodes (genes), 79,159 edges (interactions), and 17 primary modules (Fig. 2A–C, Table 1). Of note, 10 out of 17 modules (537 genes) were downregulated, whereas only 214 genes distributed in 7 modules were overexpressed in iPSCs (Table 1). Based on Gene Ontology (GO, www.geneontology.org), these modules primarily function in transcription (M2, M11, M12, and M16), development (M6, M7, M13, and M17), immune response (M15), metabolism (M4, M5, and M14), and enzyme activities for broad bioprocesses primarily including metabolism (M3, M8, M9, M10) (Table 1). We thereafter refer these primary functional modules as metabolism (transcription, development, immune response, and metabolism).
**MODULES IN iPSCs AND ESCs**

![Diagram of 17 modules and gene networks](image)

**FIG. 2.** Gene networks are differentially expressed in human iPSCs and ESCs. (A) An overview of 17 network modules identified by weighted gene co-expression network analysis (Table 1 for detail). (B) An example of a module (light green) shows network component connections. Node color denotes differential expression level (iPSCs/ES), green for down-regulation, and red for up-regulation. Node size represents the importance of a node; bigger size indicates more importance. Edge denotes interaction strength, thicker for stronger interactions. (C) A holistic view of all 17 modules. Ten and 7 out of total 17 modules are downregulated (green nodes) and upregulated (red nodes) in iPSCs, respectively. The same illustration strategy was used in all network figures in this study. Color images available online at www.liebertonline.com/scd

**Genes of network modules are coherently co-expressed and can be used as variables to distinguish iPSCs from ESCs**

Genes in our 17 identified differentially expressed modules are consistently co-expressed (Fig. 3A) and coherent (Fig. 3B). To determine whether these modules can be used as variables to distinguish iPSCs and ESCs, we added 3 independent datasets (Supplementary Table S1; Supplementary Data are available online at www.liebertonline.com/scd; Materials and Methods section) and quantified these modules by calculating the module eigengene (see definition in Materials and Methods section). The quantitative values were used to train discriminative models and to predict the accuracy of our models in discriminating iPSCs from ESCs. SVMs [28] were employed here for classifying samples and the accuracy was measured by calculating both correct proportion and kappa value (Materials and Methods section). Based on the 17 modules (Table 1, Fig. 1) and 4 meta-modules (transcription, metabolism, immune response, and development), we used re-iterative random sampling (of 3,000 times) on 70% samples as training set and the rest as
FIG. 3. Differential activations of functional modules in iPSCs and ESCs are inversely correlated with DNA methylation and can be used for annotation of iPSCs and ESCs. (A) We use the heatmap of M2, blue module (Table 1) as an example to show gene co-expression in iPSCs (blue) and ESCs (red). In the heatmap, each row represents a gene, and each column denotes a sample. Red and blue represent up- and downregulated genes, respectively. (B) Gene expression in a module is all differentially regulated in the same pattern across all observed conditions (i.e., coherently expressed). (C–F) Predictive model using SVMs. Accuracy represents the mean of 3,000 random samplings for every possible module permutation with sample size from 1 to 17 modules (C) or 1 to 4 meta-modules (D). (E) SVM plot visualizing the classification of iPSCs and ESCs based on 2 meta-modules, transcription and metabolism. “X” denotes support vectors and “O” with corresponding color represents classified true groups (iPSCs/ESCs). The colored background visualizes the predicted group regions. (F) Density dot plot showing overall inverse correlation between gene expression and DNA methylation based on all genes in total 17 modules. Darker blue indicates higher density of genes. SVMs, support vector machines. Color images available online at www.liebertonline.com/scd
MODULES IN iPSCs AND ESCs

![Network module membership in subtypes of iPSCs.](image)

**FIG. 4.** Network module membership in subtypes of iPSCs. Correlation of total network module membership (x-axis) and module membership of specific subtypes of iPSCs (y-axis). Three subtypes of iPSCs were presented here, factor-free iPSCs (rho=0.85), iPSCs with nonintegrating episomal vectors (rho=0.92), and retrovirus-integrating iPSCs (rho=0.78).

testing dataset, and found that the overall accuracy reached 96% (kappa 0.90) and 95.9% (kappa 0.90), respectively (Fig. 3C, D). Even with 2 meta-modules, transcription and metabolism, iPSCs and ESCs can be classified with an accuracy of 94% (kappa 0.85) (Fig. 3E). This indicated that the modules identified in this study can be used as quantitative variables to discriminate these 2 cell types.

The expression level of modular genes is inversely correlated with DNA methylation

To explore the mechanisms underlying the functional module differences between iPSCs and ESCs, we compared genome-wide DNA methylation profiling of iPSCs and ESCs by performing an independent microarray experiment on 5 iPSCs and 5 ESCs (Materials and Methods section). If DNA methylation plays a role in regulating gene expression, we expect genes with lower expression to have higher levels of methylation at their promoters. After correlating methylation profiles with gene expression data used for building the network (751 genes total), we found an overall inverse correlation between gene expression and DNA methylation in the network (Fig. 3F). However, a large fraction of genes do not show methylation changes (middle in Fig. 3F), indicating that DNA methylation only partially accounts for the gene expression differences between iPSCs and ESCs.

The network modules are conserved across in various iPSCs

To investigate the conservation of network modules across different types of iPSCs, we examined the network module membership of 3 types of iPSCs (ie, virus-integrating-iPSCs, factor-free iPSCs through the cre/loxP system, and vector-free iPSCs with episomal-vectors) by measuring the network module eigengene-based connectivity [26] (Materials and Methods section). We first calculated the total network module membership and then calculated the network module membership of each type of iPSCs separately. We then correlated the total network module membership with each type of iPSCs (Materials and Methods section). The module memberships of these 3 types of iPSCs are highly correlated with total modules (rho 0.76-0.92; Fig. 4). The slight difference in correlation coefficient between cell types from 0.92, 0.85, to 0.78 may result from sample sizes in different subsets of above 3 types of iPSCs and biological experiment variations, but overall they are very similarly correlated to total network. Therefore, our data indicate that the network modules identified above are overall conserved in different types of iPSCs.

To determine the extent of functional module conservation in different species, we investigated the conservation of these functional modules in mouse iPSCs. Global analysis of mouse iPSC and ESC datasets available in GEO database from different microarray platforms revealed that the primary human functional modules, including transcription, development, metabolism, and immune response, are consistently conserved in all mouse datasets (Fig. 5, Supplementary Fig. S1A-C). Together, our data demonstrate that distinctively expressed functional modules are highly conserved molecular features distinguishing iPSCs and ESCs.

Analysis of hub genes in the network

The highly connected hub genes may play crucial roles in a network; we searched for such hub genes by ranking the genes based on the network connectivity that considers both the network topology and the eigengene-based connectivity (Materials and Methods section) [26,27]. We selected the top 35 hub genes (Table 2, Supplementary Table S2), which represents ~5% of total genes in this network, including the eukaryotic translation initiation factor 2-alpha kinase 1 (EIF2AK1, P value <2.0e-19). In silico knockout [33] of these top genes resulted in significant perturbations in network diameter compared with random simulation (Supplementary
FIG. 5. Different functional modules are conserved in mouse iPSCs and ESCs. Gene ontology analysis showed functional modules are conserved in mouse iPSCs versus ESCs comparisons, similar to human iPSCs and ESCs. Presented here are the functional module annotations of mouse iPSCs versus ESCs extracted from GEO number GSE16082 as measured by Illumina microarray platform. Larger node size represents the higher density of genes, and dark node color represents greater significance (adjusted *P* value < 0.05). For illustration purposes, we grouped all modules with similar functions into a meta-module and labeled it with their corresponding functions. Similar functional conservations were observed in Affymetrix mouse datasets (Supplementary Fig. S1). Color images available online at www.liebertonline.com/scd

Fig. S2A), further indicating that these genes contribute significantly to the network structure. These key genes are found in large modules with high connectivity, such as module M1, M3, M6, and M16 (Supplementary Fig. S2B), and exhibit highly coherent expression with their modules (Fig. 6A), indicating that they are central genes within their modules. Surprisingly, the top genes show similar expression across different iPSCs (Fig. 6B–D), indicating that they are consistently important for all iPSCs.

While relatively little is known about most of these hub genes, a few genes with known functions can be classified into the following functional groups: cell development and differentiation (HMGB3, RORB), immune response and translation initiation (EIF2AK1 and ABHD2), transcription (TCEB3), metabolism (GRIN2D), magnesium ion binding and enzyme activity (RPS6KA2, GRIN2D, B4GALT6), and calcium-dependent phospholipid binding (ANXA11). This indicated that functional differences for these 2 cell types are still primarily in transcription, development, metabolism, and immune response, consistent with our finding observed above. Thus, our data uncovered top genes that regulate their corresponding protein module expression and potentially contribute to functional differences between iPSCs and ESCs.

The largest and most significantly altered module primarily functions in transcription

After viewing the global properties of the entire network, we next examined details of particular modules. We first
The whole module (M2, blue) primarily functions in mediating gene expression. We next decomposed the most predominant module (M16, turquoise; Table 1 and Supplementary Fig. S4) with 163 genes and 12,692 interactions, downregulated in iPSCs. Because most genes in the module have unknown functions, we focus on genes with known functions to discuss the primary functions of this module. The majority of key proteins, 17 of 35 crucial genes identified in the network, are located in this module (Fig. 8A). These genes strongly interact with each other (Fig. 8A) and it is therefore difficult to determine the most important gene. Key genes with known functions are associated with transcription (including RORB and TCEB3), indicating that transcription is the primary process mediated by the key genes in the entire network differentially expressed between iPSCs and ESCs.

Two other primary functional protein complexes were found in the modules DNA repair (Fig. 8B) and immune response (Fig. 8C). The DNA repair complex consists of 6 genes centered around EXO1 and RECO1 (Fig. 8B), including EXO1, FLJ33220, RECQ5, WDR51L1, EME2, and FAH, whereas immune response complex contains 9 genes centered around EXO1 and TREC1, including EXO1, IGHC1, CDC42, IL17A, LST1, ANXA11, TIRAP, TREC1, and TCF12. These 2 groups overlapped very well both in interactions and in key genes like EXO1 (Fig. 8D), suggesting that the overall function of these 2 complexes is immune response to DNA damage. Together, our module data suggest that the differences in transcription and immune response between iPSCs and ESCs are primary molecular features distinguishing these 2 cell types.

Discussion

A critical question we must answer before applying iPSCs in regenerative medicine is how close iPSCs resemble ESCs and whether there are any features distinguishing them. Here, we reveal that iPSCs generated to date still inherently express distinctive transcriptome compared with ESCs, and that these 2 cell types can be distinguished by several basic biological modules.

Experimental conditions such as cell culture, cell handling, and treatment conditions have been proposed as factors that contribute to stochastic variations in iPSCs transcriptome [1-2,10]. This seems true after observing lab-specific iPSC transcriptome profiling [11,12]. However, these lab-specific patterns were drawn from microarray analyses without adjusting for batch effects, which is notorious for misleading microarray data interpretation [15]. In addition, these patterns [11,12] were generated from cluster analysis that has low sensitivity for discriminating samples with high dimensions. In this study, we removed the batch effects from all datasets and employed between-group analysis [32] to re-analyze the iPSC samples. Between-group analysis uses a standard conversion method such as correspondence analysis to calculate an ordination of sample groups rather than that of individual microarray samples and thus it has a discriminating power compatible to artificial neural network with high sensitivity. Our analysis revealed that the lab-specific iPSC profiling is a consequence of batch effects in microarray data (Fig. 1A, C right panel) and that, after
removing batch effects, we find iPSCs are clearly separated from ESCs (Fig. 1B, C right panel). This indicates that human iPSCs inherently express distinctive transcriptome compared with ESCs.

Here, we employed systems biology approaches based on WGCNA to systematically investigate the system-wide biological picture between these 2 cell types and revealed conserved molecular features distinguishing these 2 cell types. Our analysis revealed a network containing 17 modules differentially expressed in iPSCs and ESCs (Fig. 2, Table 1). These modules can be grouped into meta-modules based on functions and they primarily function in transcription, metabolism, development, and immune response. Strikingly, the functional modules are highly conserved in various iPSCs (Fig. 4). This conservation relationship was measured by the module membership correlation based on the network eigengene scores, which uses the principle component of high dimension data and thus captures the maximum information that may explain the natural relationship of the variables.

The modules identified in this study can be used as quantitative variables to classify samples and to predict the new samples (Fig. 3). By employing SVMs, our module-based models successfully discriminate these 2 cell types with a very high accuracy, ~96% for models based on both 17 modules and 4 meta-modules (transcription, metabolism, immune response, and development). Even with 2 meta-modules (transcription and metabolism), our model reaches a 94% accuracy (Fig. 3C–E). Together, coherent co-expression, conservation, and discriminating powers of these modules suggest that these functional modules identified here serve as inherently conserved features distinguishing iPSCs and ESCs. This further suggests that these 2 cell types exhibit the distinctive differences in fundamental biological
functions such as in transcription and metabolism. Consistently, recent studies have observed improvements in transcription and metabolism during iPSC production by adjusting transcription factor composition and hypoxia condition [34], adding microRNAs [35], and other factors like vitamin D [36]. Furthermore, enzyme activity differences in metabolism between iPSCs and ESCs may explain the recent observations showing that modified culture medium enhances the iPSCs generation [36]. Therefore, iPSCs have unique distinguishing features to ESCs.

Altered expression of functional modules may be modulated by many mechanisms, including epigenetic and genetic factors. Our data uncovered an overall inverse correlation between module expression and DNA methylation level (Fig. 3). We observed a similar trend even when we expanded our data set with 67 samples (unpublished data), suggesting that DNA methylation may serve as one epigenetic mechanism underlying functional module differences. Our present result on DNA methylation differences parallels the most current observations showing that iPSCs retain DNA methylation
patterns from original somatic cells [10,37], and that iPSCs differentially express a panel of DNA methylation sites compared with ESCs [38–40]. Further biological experiments and bioinformatics algorithms are needed to fully understand the role of DNA methylation in regulating these modules. Recently, copy number variations are uncovered in iPSC compared with the parental somatic cells, suggesting that genetic changes can also take place in iPSC derivation [41,42]. Thus, we cannot rule out that genetic changes may also contribute to functional differences of human iPSCs and ESCs.

Our study systematically reveals inherent functional modules that are uniquely activated in iPSCs. Our findings provide an avenue to guide the further efforts on overcoming the barriers of transcriptional differences between iPSCs and ESCs.

Acknowledgment

The authors deeply appreciate Dr. Peter Langfelder for providing assistance during data analysis. This work is supported by CIRM RC 1-00111 grant, NIH PO1 GM 081621, 2011CB965102, and 2011CB966204 from Ministry of Science and Technology in China.

Author Disclosure Statement

No competing financial interests exist.

References

high-throughput characterization of pluripotent cell lines. Cell 144:439-452.


WANG ET AL.

Address correspondence to:
Dr. Guoping Fan
Department of Human Genetics
David Geffen School of Medicine
UCLA
Los Angeles, CA 90095

E-mail: gfan@mednet.ucla.edu

Received for publication December 17, 2010
Accepted after revision May 03, 2011

Prepublished on Liebert Instant Online XXXX XX, XXXX
Chapter 4
Molecular signature of primary retinal pigment epithelium and stem-cell-derived RPE cells

Jo-Ling Liao¹, Juehua Yu¹, Kevin Huang¹, Jane Hu², Tanja Diemer², Zhicheng Ma¹, Tamar Dvash¹, Xian-Jie Yang², Gabriel H. Travis², David S. Williams², Dean Bok² and Guoping Fan¹,*

¹Department of Human Genetics and ²Jules Stein Eye Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

Received May 26, 2010; Revised July 25, 2010; Accepted August 9, 2010

Age-related macular degeneration (AMD) is characterized by the loss or dysfunction of retinal pigment epithelium (RPE) and is the most common cause of vision loss among the elderly. Stem-cell-based strategies, using human embryonic stem cells (hESCs) or human-induced pluripotent stem cells (hiPSCs), may provide an abundant donor source for generating RPE cells in cell replacement therapies. Despite a significant amount of research on deriving functional RPE cells from various stem cell sources, it is still unclear whether stem-cell-derived RPE cells fully mimic primary RPE cells. In this report, we demonstrate that functional RPE cells can be derived from multiple lines of hESCs and hiPSCs with varying efficiencies. Stem-cell-derived RPE cells exhibit cobblestone-like morphology, transcripts, proteins and phagocytic function similar to human fetal RPE (fRPE) cells. In addition, we performed global gene expression profiling of stem-cell-derived RPE cells, native and cultured fRPE cells, undifferentiated hESCs and fibroblasts to determine the differentiation state of stem-cell-derived RPE cells. Our data indicate that hESC-derived RPE cells closely resemble human fRPE cells, whereas hiPSC-derived RPE cells are in a unique differentiation state. Furthermore, we identified a set of 87 signature genes that are unique to human fRPE and a majority of these signature genes are shared by stem-cell-derived RPE cells. These results establish a panel of molecular markers for evaluating the fidelity of human pluripotent stem cell to RPE conversion. This study contributes to our understanding of the utility of hESC/hiPSC-derived RPE in AMD therapy.

INTRODUCTION

Age-related macular degeneration (AMD) is a severe retinal disease that significantly impairs vision. In the western world, AMD is the leading cause of blindness among the elderly, affecting over 30 million people worldwide (1). AMD patients are usually afflicted with degenerated and/or dysfunctional retinal pigment epithelium (RPE), which normally plays various central roles in maintaining retinal integrity and viability (2). In particular, RPE is involved in the formation of the blood-retinal barrier, absorption of stray light, supplying of nutrients to the neural retina, regeneration of visual pigment, as well as the uptake and recycling of the outer segments of photoreceptors. Consequently, loss of RPE leads to photoreceptor depletion and irreversible blindness (3). Current treatments for AMD are severely limited. Palliative treatment options are only available for the less prevalent, ‘wet’ form of the disease, including the use of anti-neovascular agents, photodynamic therapy and thermal laser therapy. However, there are no current treatments for the more widespread, ‘dry’ AMD except for the use of antioxidants to delay disease progression in the eye. Despite current treatments, patients with ‘dry’ AMD generally show poor prognosis and eventual loss of vision (4).

Cell therapy holds tremendous promise in treating AMD; directly replenishing the degenerated RPE can restore retinal function and rescue vision in AMD patients. Autologous RPE/choroid transplant attempts from periphery to central retina have demonstrated partial restoration of vision in AMD patients (5). However, autologous transplantation is limited by the scarcity and genetic predisposition to AMD of the cell source, which may affect the efficacy of transplantation (5). Pluripotent stem cells have been proposed to be

*To whom correspondence should be addressed at: Department of Human Genetics, David Geffen School of Medicine, UCLA, 695 Charles Young Drive South, Los Angeles, CA 90095, USA. Tel: +1 3102670439; Fax: +1 3107945446; Email: gfan@mednet.ucla.edu

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
an attractive alternative cell source for transplantation. Human embryonic stem cells (hESCs) can indefinitely self-renew and differentiate into any cell type found in the adult body, making hESCs a promising candidate for generating an unlimited donor source for RPE transplantation (6). In addition, recent derivation of human-induced pluripotent stem cells (hiPSCs) by forced expression of four transcription factors (Oct4, Sox2, c-myc, Klf4) in fibroblasts has created an additional cell source for cell therapy (7). Various studies report that hiPSCs closely resemble hESCs and have been proposed to be promising surrogates for hESCs (7–9). hiPSCs have the added advantage of avoiding immunological complications and ethical controversies that are typically associated with handling hESCs (10). In addition, hiPSCs have the potential to become a platform for personalized medicine by allowing a patient’s own cells to become a source for therapeutic tissue (11).

Previous studies on differentiating RPE cells from stem cells demonstrate that stem-cell-derived RPE cells have molecular characteristics similar to primary RPE cells (2,12,13). In addition, the transplantation of stem-cell-derived RPE can partially restore visual function in the retinal dystrophy rat model (12,14,15). However, despite a significant amount of research on the derivation of functional RPE cells from various stem cell sources, no systemic comparison has been done between these stem-cell-derived RPE cells and primary RPE cells. In order to realize the therapeutic potential of stem-cell-derived RPE cells, it is important to ensure that stem-cell-derived RPE cells can recapitulate both functional and genetic characteristics of primary RPE cells.

RESULTS

**Differentiation and expansion of putative RPE cells from hESCs and hiPSCs**

To determine the ability of various lines of hESCs and hiPSCs to differentiate into RPE cells, we followed a previously described differentiation protocol using a total of 11 cell lines (Supplementary Material, Table S1) (12). Pigmented cells spontaneously arise from differentiating hESCs and hiPSCs after 3–4 weeks of culture in bFGF-free hESC culture media. Pigmented clusters grew in size and number after an additional 2–3 weeks of culture. Although all cell lines were able to generate pigmented clusters reproducibly, various lines of hESCs and hiPSCs displayed varying differentiation efficiencies. H9 and H1 lines showed the highest efficiencies, giving rise to over 200 pigmented clusters per cell culture dish (Fig. 1A). HUES12, HUES15 and HSF1 cell lines showed fairly similar efficiencies with about 150 pigmented clusters. HSF6 cell lines showed relatively lower efficiencies giving rise to less than 100 pigmented clusters. All lines of hiPSCs gave rise to approximately 150 pigmented clusters, despite different parental cell types. hiPSC cell lines derived from neuroectodermal and endodermal lineage both generated pigmented clusters at similar efficiencies.

Stem-cell-derived pigmented clusters were manually excised and expanded to monolayers on Matrigel. These stem-cell-derived pigmented sheets appeared homogenous in the inner part of a cluster; however, cells on the edges generally showed less pigment, suggesting that they may be in a less differentiated state. This observation is consistent with previous reports that stem-cell-derived RPE cells expand by
Figure 2. Stem-cell-derived RPE exhibit characteristic marker genes of RPE. (A) and (B) are immunofluorescent labeling of RPE65 and ZO1 in hESC-RPE, respectively. (C) RT–PCR analysis of human RPE gene markers in different lines of hESC-RPE and hiPSC-RPE. (D) Western blot analysis of RPE65 and GAPDH proteins in hESC-RPE, hiPSC-RPE and iRPE. Human RPE cells cultured for 3 months in vitro were used as a positive control in this assay. Scale bar: (A) and (B), 50 μm.

de-differentiation, proliferation and then re-differentiate to acquire pigmentation (12). The stem-cell-derived pigmented sheets were cultured for 2–3 months and characterized through various molecular and functional assays. We pursued and characterized the lines that could be most efficiently differentiated (H9-RPE, HSF1-RPE, HiPS2-RPE, HiPS18-RPE and NPC2-IPS#3-RPE). HESC-RPE and hiPSC-RPE expanded sheets exhibited polygonal, cobblestone-like morphology under bright field microscopy, consistent with normal fetal RPE (iRPE) cell morphology (Fig. 1). Electron micrograph results revealed that hESC-RPE have apical microvilli, basal end-feet and melanin structures similar to human RPE (Supplementary Material, Fig. S1A). We observed a small number of flattened cells scattered beneath the hESC-RPE monolayer. These non-RPE cells resemble fibroblasts and are embedded in Matrigel substrate. Overall, consistent with previous reports, our results indicate that multiple lines of hESCs and hiPSCs can spontaneously differentiate to clustered pigmented cells, which further differentiate into monolayer RPE-like cells in defined culture environments.

Stem-cell-derived RPE cells express genes and proteins associated with primary RPE cells

To further determine whether hESC-RPE and hiPSC-RPE mimic normal RPE cells, we asked if these stem-cell-derived RPE cells express the appropriate RPE markers. Using reverse transcription–polymerase chain reaction (RT–PCR), we assayed for the expression of several known genes that are expressed in human RPE cells. We first examined the regulatory gene transcripts central to RPE development, including the early eye field development marker, Pax6 and RPE-specific transcription factors MITF and OTX2. Both MITF and OTX2 are necessary for RPE differentiation; in particular, MITF is known to be associated with the onset and maintenance of pigmentation and is an established marker of RPE during the course of eye development. We also examined the expression of RPE functional markers in stem-cell-derived cells. These markers include the secreted (PEDF), membrane-associated proteins (BEST1, ZO1) and most importantly, the visual cycle proteins (RPE65, LRAT, CRALBP). For comparison, 3 months in vitro, cultured human RPE cells (16) were used as a positive control in all assays.

Interestingly, RT–PCR detected Pax6 in hESC-RPE and RPE but not in hiPSC-RPE. MITF and OTX2 were both detected in hESC-RPE and iRPE, whereas hiPSC-RPE expressed only MITF but not OTX2. Secreted and membrane-bound protein such as PEDF, BEST1 and ZO1 were seen in all stem-cell-derived RPE cells. Furthermore, RPE65, LRAT and CRALBP transcripts were detected in both types of stem-cell-derived RPEs (Fig. 2C). In addition, RPE65 protein expression was confirmed in both hiPSC-RPE and hESC-RPE by immunoblot (Fig. 2D). These results indicate that both hESC-RPE and hiPSC-RPE exhibit many functional markers characteristic of primary iRPE cells. Moreover, hESC-RPE more closely resembles iRPE than hiPSC-RPE, whereas hiPSC-RPE appears to be in a unique differentiation state.
To determine whether hESC-RPE has a definitive phagocytic function, we first examined microarray data and found that expression of phagocytic genes (MERTK, LAMP2, VPD and GULP1) are comparable between hESC-RPE and fRPE (see below). Second, using a previously characterized cell culture assay, we compared the ability of the hESC-RPE cells to bind, ingest and digest mouse rod outer segments (ROSs) disk membranes with that of human fRPE cells. After 30 min incubation with isolated ROSs, the hESC-RPE cells had nearly twice as many ROSs bound to their surface, although the number of ingested ROSSs was comparable to fRPE cells (Fig. 3). The rate of digestion was determined by comparing the number of ROSs remaining in the cells after subsequent 2 h incubation in the absence of ROSs. The calculated proportion of ROSs digested during this period was similar in the hESC-RPE and fRPE cells (Fig. 3). The hESC-RPE cells therefore appeared comparable to fRPE cells in their ability to ingest and digest ROSs. There was, however, a difference in the number of ROSSs that bound to the cell surface. Even after the 2 h incubation in the absence of ROSs, hESC-RPE had many more ROSSs bound to their surface (~20 ROSSs/field of view), suggesting that they may express more ROS receptor that is not involved with ingestion.

Transcriptome analysis of fRPE, hESC-derived RPE and hiPSC-derived RPE

In order to gain a more comprehensive understanding of the differentiation state of stem-cell-derived cells, we generated global gene expression profiles of stem-cell-derived RPE cells and compared them with native fRPE, primary culture fRPE cells, undifferentiated hESCs and fibroblasts. Probe intensity signals were quantile-normalized and log-transformed across samples using the limma package in Bioconductor. To increase the statistical power of our analysis, we incorporated microarray data sets of various hESCs and somatic tissues including brain, liver, lung and melanocytes from the Gene Expression Omnibus (GEO) database (Supplementary Material, Table S2).

To determine the degree of similarity or difference between stem-cell-derived RPE cells and other cells and tissues, we first performed hierarchical clustering and principal component analysis (PCA). Hierarchical cluster and PCA both demonstrate distinct separation of undifferentiated hESCs, somatic tissues and native/cultured fRPE clusters. As expected, both methods show technical replicates of somatic cells cluster closer together in comparison with clusters of similar biological samples (Fig. 4). The dendrogram shows that hESC-RPE clustered together with fRPE whereas hiPSC-RPE clustered away from the fRPE. Importantly, stem-cell-derived RPE cells do not cluster together with human melanocytes, which is another pigmented cell type derived from neural crest. For PCA analysis, we focussed on the first three principal components, which captured over 70% of the total variance. Consistent with hierarchical clustering results, biplots between the first three principal components revealed close clustering of stem-cell-derived RPE cells and fRPE while clustering away from both hESCs and somatic cells (Fig. 4C and D).

Furthermore, pair-wise comparisons between gene expression profiles show more similarity between hESC-RPE and fRPE than between hiPSC-RPE and fRPE (Fig. 4A). This is further supported by a higher Pearson correlation between hESC-RPE and fRPE cells. In addition, both lines of hESC-RPE are well correlated with each other (Pearson’s correlation = 0.948). The comparison of hESC-RPE and fRPE data sets revealed 88 transcripts that were differentially expressed at least 1.5-fold, 39 of which were up-regulated and 49 were down-regulated. Gene ontology revealed that up-regulated genes in hESC-RPE correspond to biological processes such as cell proliferation and neuronal differentiation, whereas down-regulated genes were associated with visual perception and gas transport. The comparison of hiPSC-RPE and fRPE data revealed 224 transcripts that were differentially expressed, 30 of which were up-regulated and 194 were down-regulated. The down-regulated genes in hiPSC-RPE, similar to hESC-RPE, are involved in gas transport and visual perception. Interestingly, an analysis of over-expressed transcripts in hiPSC-RPE revealed that the majority of these genes are involved in epithelial development and the inflammatory response. Notably, gene network analysis of
multiple lines of hESCs and hiPSCs also identified many immune response genes that are significantly differentially expressed between hESCs and hiPSCs (unpublished data). This suggests that hiPSC-RPE may retain the gene signatures of hiPSCs. Overall, these data indicate that hESC-RPE is reproducibly more closely related to fRPE than hiPSC-RPE.

**Figure 4.** Microarray analysis of genome-wide gene expression of primary fRPE and stem-cell-derived RPE cells. (A) Pair-wise comparisons between fRPE, hESCs and stem-cell-derived RPE cells. Linear regression lines are shown in red and the green curves represent $P = 0.05$. Data points outside the green curve represent genes that have significant, differential expression in each cell type. Numbers above the green curve indicate the total number of genes that are up-regulated in the y-axis, and numbers below the green curve represent genes up-regulated on the x-axis. Pearson’s correlation coefficients are shown for comparison of each data set. (B) Unbiased hierarchical clustering showing hESC-RPE cluster closer to fRPE than to hiPSC-RPE. Various somatic tissues and additional hESCs from GEO database were included in gene expression analysis to increase statistical power. (C and D) Biplots of three predominant principal components (PC1, PC2 and PC3) demonstrate that samples cluster to three distinct groups of somatic tissues, hESCs and RPE.

**Stem-cell-derived RPE cells express low levels of marker genes associated with aging and AMD**

We examined genes that are involved in RPE differentiation, pigment synthesis, phagocytic activity and vitamin A metabolism. HESC-RPE showed comparable expression levels with fRPE for the majority of these genes (Fig. 5A). We
Figure 5. Comparative analysis of gene expression levels for pluripotency and RPE-related genes in hESCs, hESC-RPE, hiPSC-RPE and fRPE. Normalized expression levels were obtained from Agilent whole-genome microarrays. (A) Genes that are significantly expressed in pluripotent stem cells and RPE-associated markers were selected. RPE-associated markers are expressed in hESC-RPE, many at levels close to human RPE cells. Two technical replicates were used for all cell types. (B) Expression of genes associated with aging retina and complement cascades.

next examined whether stem-cell-derived RPE cells express genes involved in AMD pathogenesis. AMD is considered a complex genetic disease involving interactions between multiple genes and environmental risk factors, with age and family history being the two strongest risk factors. In addition, recent genetic studies have shown strong association between single nucleotide polymorphism (SNP) variants in complement cascade, components of the innate immune system and AMD pathogenesis (17). We examined the expression of genes which were previously identified to be differentially expressed in young and aging human retina (18), including genes involved in energy metabolism (CKB, PGAM1, ENO3) and stress response (CIRBP, COL7A1, MACS). We observed a low level of expression for both metabolic and stress response genes in stem-cell-derived RPE cells and fRPE. These results indicate that stem-cell-derived RPE cells, though subjected to culture conditions, showed similar energy metabolism and stress response as normal fRPE.

Furthermore, we looked at expression of common genes involved in the complement cascade. A number of complement system protein variants and complement regulatory protein variants are shown to be strongly associated with AMD (17,19). These include complement factor H, complement factor B, complement component 2 (C2) and complement component 3 (C3). Stem-cell-derived RPE showed comparable expression levels to fRPE in all of the complement factor genes we examined (Fig. 5B). Overall, these results indicate that stem-cell-derived RPE cells express a normal level of many RPE-associated markers as well as genes associated with complement regulation.

Human RPE have distinct signature genes from stem-cell-derived RPE cells

A recent comprehensive study comparing gene expression profiles of fetal and adult RPE with somatic tissues identified 154 signature genes that are unique to RPE (20). We asked if these RPE signature genes can be used to classify fRPE, stem-cell-derived RPE and hESCs. Our cross-reference analysis showed that 151 candidate signature genes in their Affymetrix platform can be identified in our Agilent microarray. Among these 151 signature genes, we found that 43 genes were not significant in the differential expression between fRPE and hESCs (Supplementary Material, Table S3). Gene ontology analysis revealed that these genes are associated with protein transport and fatty acid metabolism, suggesting that RPE may share many metabolic properties with hESCs (Fig. 6B). Although these 43 genes are not significantly differentially expressed between hESCs, it does not exclude the possibility that these genes are important for RPE maintenance. Of the remaining 108 RPE signature genes, 87 genes showed robust expression in fRPE, suggesting that these 87 fRPE-specific genes better represent true RPE signature genes (Fig. 6A). Consistent with global gene expression comparison as seen in Fig. 4, hESC-RPE also shared more signature genes with fRPE than hiPSC-RPE. Venn diagram analysis of fRPE, hESC-RPE and hiPSC-RPE indicated 42 genes are commonly shared by these cells, suggesting that these 42 genes are the core RPE differentiation genes (Fig. 6A). Gene ontology analysis indicated that these genes are involved in various pigmentation, visual perception and eye development pathways, consistent
Figure 6. Analysis of RPE signature genes in IRPE, hESC-RPE and hiPSC-RPE. (A) The Venn diagram shows the shared/unique genes among 108 selected RPE signature genes for IRPE, hESC-RPE and hiPSC-RPE. (B) Functional annotation from gene ontology analysis of 43 IRPE signature genes that are shared with undifferentiated hESCs. (C) Twenty-one IRPE genes that are under-expressed in stem-cell-derived RPE cells. (D) Forty-two core RPE signature genes that are shared by IRPE and stem-cell-derived RPE cells.

with their role in RPE differentiation pathways (Fig. 6D). These results suggest that despite morphology and functional similarities, genetic profiles of stem-cell-derived RPE cells are moderately different from normal IRPE cells.

Notably, by cross-comparison of the 87 RPE signature genes in IRPE with gene expression in hESC-RPE and hiPSC-RPE cells, we found that 21 signature genes that are exclusive to IRPE, but not shared with stem-cell-derived RPE cells (Fig. 6A, Supplementary Material, Table S5). Gene ontology analysis showed that these genes are involved in various eye development pathways such as BMP signaling, lens development and eye and lens morphogenesis (Fig. 6C). This suggests that stem-cell-derived RPE may lack robust expression of some critical genes that are involved in eye development.
DISCUSSION

The RPE is critical for retinal viability and function and has been the focus for AMD and its therapeutic interventions. Although a significant amount of research has focussed on deriving functional RPE cells from various stem cell sources, little is known about the global transcriptional profiles of these stem-cell-derived RPE cells. In this study, we have derived functional RPE cells from various lines of hESCs and hiPSCs and generated global expression profiles of stem-cell-derived RPE and human IRPE cells. We found that hESC-RPE resembles IRPE more closely than hiPSC-RPE. The expression profiles of hiPSC-RPE suggest that they are in a relatively immature differentiation state. Also, we identified a set of genes that are exclusively expressed in human IRPE. This set of genes may serve as reliable molecular signatures of IRPE and provide future standards for scoring the differentiation state of stem-cell-derived RPE cells. Together, these results offer critical insights into the therapeutic use of stem-cell-derived RPE cells in treating AMD.

Stem-cell-derived RPE cells have a highly similar morphology to human IRPE. These cells arise spontaneously from stem cell sources 3–4 weeks after differentiation, consistent with the differentiation timeline shown by other groups. We observed varying differentiation propensities between cell lines. This may be a result of genetic background differences between each cell line, especially in hiPSCs. Previous reports have shown some cell types are more efficiently reprogrammed (7,21,22); it is possible that these biases are persistent during differentiation to RPE cells.

hESC-RPE and hiPSC-RPE express a panel of RPE gene transcripts similar to cultured IRPE cells. Co-expression of the early RPE development marker, PAX6, with various mature markers suggests that hESC-RPE has some heterogeneity, which may be due to immature cells around the leading edges of hESC-RPE expanding sheets. hESC-RPE were previously shown to expand through a de-differentiation mechanism, which involves pigmented cells showing de-pigmentation, then entering into the active cell cycle before gaining pigmentation again (12). Interestingly, a majority of the RPE-associated markers was detected in hiPSC-RPE, but not the early marker, PAX6, and the late marker, LRRAT, suggesting that hiPSC-RPE may be in a unique differentiation state. Recent reports have shown some loci are more resistant to reprogramming in hiPSCs, which may affect the expression of certain genes (9,23,24). There appear to be several discrepancies between hESC-RPE, hiPSC-RPE and primary RPE cells, but it remains unclear whether these differences will affect the utility of stem-cell-derived RPE in the clinical setting.

Global gene expression analysis demonstrated that hESC-RPE resembles IRPE more closely than hiPSC-RPE. Pair-wise comparisons between hESC-RPE to IRPE and hiPSC-RPE to IRPE revealed that stem-cell-derived RPE cells show under-expression of genes involved in visual perception. Interestingly, some of the up-regulated transcripts in hiPSC-RPE are involved in immune responses. This observation is consistent with transcriptional and methylation profile comparisons between hESCs and hiPSCs (unpublished data), suggesting hiPSC-RPE may retain specific reprogramming gene signatures.

Furthermore, we found similar gene expression levels of complement associated proteins previously reported to be associated with AMD in stem-cell-derived RPE and IRPE cells. Allele variants for different complement cascade-associated proteins have been implicated as major risk factors for the predisposition of AMD (25,26). Further analysis of SNPs in each stem cell line is required to determine whether they carry AMD risk variants. Our analysis of aging retina-associated genes shows that cultured stem-cell RPE mimic IRPE and do not show signs of aging. This addresses the concern that extended culture may induce the expression of age-related genes and suggests that culture-derived RPE may be suitable replacements for IRPEs in cell transplantation therapy.

Recent microarray-based studies have identified RPE-specific genes by comparing human RPE with various retinal and somatic tissues (20,27). However, comparisons between two separate studies showed very few RPE signature genes actually overlapped. These discrepancies may arise from differences in the experimental design and tissues used in their respective analyses. We chose to cross-reference our data with Miller’s data set due to the more comprehensive nature of their study (20). Using their identified signature gene list, we compared their expression with hESCs and refined the list by excluding genes that are not significantly differentially expressed with hESCs. Surprisingly, a small group of eye developmental genes are uniquely expressed in IRPE but not in stem-cell-derived RPE cells, indicating that several eye development pathways remain underdeveloped in stem-cell-derived RPE cells. It is possible that using different differentiation protocols may yield activation of these pathways in stem-cell-derived RPE cells. Previous studies using directed differentiation to recapitulate in vivo development of RPE first directed differentiation toward the neuronal lineage, followed by differentiation toward the RPE fate. It will be important to examine the differences between spontaneous differentiation and directed differentiation and find optimal culture conditions for deriving RPE cells that best resemble native RPE.

Our results highlight significant differences between expression profiles of stem-cell-derived RPE cells and IRPE. To fully realize the potential of stem-cell-derived RPE in cell-based therapy, future work using in vivo models will elucidate the utility of stem-cell-derived RPE in restoring vision. This study establishes a standard for expanded analyses of expression profiles in additional cell lines. Our findings represent an important step toward optimizing the future application of stem-cell-derived RPE for transplantation into AMD patients.

MATERIALS AND METHODS

Pluripotent stem cell culture

The hESC lines h9 and HSF1 were maintained as described previously (28). The induced pluripotent stem cell lines HiPSC2, HiPSC1 and NPC2-iPS#3 were maintained under similar conditions with a 1:1 ratio of hESC medium to conditioned medium. Conditioned medium was prepared by the incubation of hESC medium overnight with MEF cells.
Differentiation, enrichment and culture of pigmented cells

Differentiation of hESCs and hiPSCs to RPE cells followed a previously described protocol with minor modifications (12). All cells were cultured in 10 cm cell culture dish (Corning) or six-well cell culture plates (Costar). HESC-RPE and hiPSC-RPE were formed when hESCs colonies were allowed to become over-confluent on an MEF density of 1 × 10^6 cm^-2. When the borders of individual hESC colonies contacted each other at approximately 7–10 days post-passage, the medium was changed daily using basic hESCs medium lacking bFGF. Pigmented foci appeared in over-confluent hESCs cultures 3–4 weeks after the use of bFGF-deficient hESC medium.

Following formation, pigmented clusters were excised manually using syringe needles under a dissecting microscope. This approach was only possible after a cluster had reached at least 1 mm in diameter. During this procedure, we avoided the surrounding, non-pigmented material prior to placement of the pigmented cluster in 24-well culture dishes coated with growth factor reduced Matrigel™ (BD Biosciences, diluted 1:30). Alternatively, Millicell-HA (Millipore) culture wells were coated with Matrigel. Two to three pigmented foci were placed in each well; hESC-RPE and hiPSC-RPE were allowed to expand on Matrigel for a further 2 months in basic bFGF deficient hESCs medium (media changes every 2–3 days). This time frame was sufficient to yield monolayer sheets of pigmented cells that could be studied further.

Fetal human RPE cell culture

The culture method for fetal human RPE cells has been described previously (16,29). Briefly, fetal human RPE cells from 16–21 weeks gestation) was collected and grown in low calcium Chex’s essential modified medium until they proliferated to confluence and released cells into the medium. These non-attached cells were collected and grown for 4–8 months on permeable Millicell-HA culture wells. The culture medium was changed to a 1:1 mixture of normal calcium CEM and Eagle’s minimum essential medium, with bovine retinal extract and 1% heat-inactivated calf serum.

Reverse transcription–polymerase chain reaction

Gene expression of stem-cell-derived RPE cells was analyzed by RT–PCR. The experimental procedure is similar to that described previously (28). Primer pairs were adopted from the published literature (12,13).

Immunocytochemistry

Immunocytochemistry was carried out as described (30). Primary antibodies used in the experiment were anti-RPE65 (Chemicon) and ZO1 (Zymed).

Western blotting

Western blotting was carried out on protein samples from cell culture dish containing pigmented sheets of putative RPE cells similar to previously described procedure (30). The primary antibody used in the experiment was RPE65 (1:5000, Chemicon) and GAPDH (1:10,000, Abcam).

ROS phagocytosis assay

The phagocytosis and digestion of mouse rod outer segments (ROSs) by hESC-derived RPE and human RPE cells, grown on filters in Millicell-HA wells, were assayed as described (31). The cultured cells were incubated with 2 × 10^6 ROSs per well for 30 min, and then washed with cold Dulbecco’s PBS to remove unbound ROSs. Some filters were incubated for a further 2 h. ROSs were labeled with the polyclonal Opn opsin antibody before and after permeabilization of the cells, in order to quantify bound and ingested ROSs. Labeling before permeabilization was followed by a secondary antibody conjugated to Alexa Fluor 488 and after permeabilization by a secondary antibody conjugated to Alexa Fluar 594 (ingested ROSs are therefore labeled only by the second antibody. ROSs that were >1 μm in diameter were counted in each field of view. At least three field of view were imaged per filter, and each treatment was performed in triplicate.

Microarray

Gene expression was analyzed by RT–PCR of total RNA extracted with TRizol reagent from cell culture dishes containing pigmented sheets of putative RPE cells. RNA was purified using RNAeasy Kit (Qiagen) following the manufacturer’s instruction. Sample labeling and microarray processing was performed as detailed in the ‘One-Color Microarray-Based Gene Expression Analysis’ protocol. The labeling reactions were performed using the Agilent Low RNA Input Linear Amplification Kit in the presence of cyanine 3-CTP (PerkinElmer Life and Analytical Sciences). Fluorescent labeled probes were purified using the Qiagen RNAeasy Mini kit (Qiagen) as described by the manufacturer. Dye incorporation was confirmed using nanodrop ND1000. For microarray hybridization, 1000 ng of cyanine 3-labeled cRNA was fragmented and hybridized on Agilent 44K Whole Human genome arrays (G4112A; Agilent Technologies) and incubated at 65°C for 17 h using the Agilent Gene Expression Hybridization Kit. The hybridized microarrays were disassembled at room temperature in Gene Expression Wash Buffer 1, then washed in Gene Expression Wash Buffer 1 at room temperature for 1 min. This was followed by a wash for 1 min in Gene Expression Wash Buffer 2 at an elevated temperature (33°C). The processed microarrays were scanned with the Agilent DNA microarray scanner immediately after washing to prevent ozone degradation and data were extracted with Agilent Feature Extraction software. Microarray data are deposited to GEO database with accession number available upon publication.

Gene expression analysis

Data analysis was carried out using both R and Matlab. R was used for importing, quantile-normalizing and log-transforming Cy3 signal intensity only. Pertinent raw data sets available from the GEO database were similarly normalized and log-transformed. To correct for batch effects, and non-biological variations between different data sets, we implemented batch
effect correction using the ComBat package in R. Hierarchical clustering was performed using Matlab’s dendrogram function. PCA analysis was performed using the princomp function and visualized using the mapplot graphical interface. Biplots for various pair-wise comparisons were also carried out in Matlab using the scatter plotting function. Differential expression analysis between RPE and ES cells was determined using a combination of t-statistic, false discovery rate, and fold change.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We thank Maren Engelhardt for assistance with phagocytosis experiment.

Conflict of Interest statement. None declared.

FUNDING

This work is supported by Translational I Grant TR1-272 from California Institute for Regenerative Medicine to X.-J.Y., G.H.T., D.S.W., D.B. and G.F. G.F. is a Carol Moss Spivak Scholar in Neuroscience.

REFERENCES

Chapter 5
Identification of novel molecular markers through transcriptomic analysis in human fetal and adult corneal endothelial cells

Yinyin Chen¹,²,³, Kevin Huang³, Martin N. Nakatsu⁴, Zhigang Xue¹,*, Sophie X. Deng⁴,* and Guoping Fan¹,³,*

¹Translational Center for Stem Cell Research, Tongji Hospital, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China, ²School of Life Science and Technology, Tongji University, Shanghai 200092, China, ³Department of Human Genetics, David Geffen School of Medicine and ⁴Cornea Division, Jules Stein Eye Institute, University of California Los Angeles, Los Angeles, CA 90095, USA

Received October 30, 2012; Revised December 4, 2012; Accepted December 11, 2012

The corneal endothelium is composed of a monolayer of corneal endothelial cells (CECs), which is essential for maintaining corneal transparency. To better characterize CECs in different developmental stages, we profiled mRNA transcriptomes in human fetal and adult corneal endothelium with the goal to identify novel molecular markers in these cells. By comparing CECs with 12 other tissue types, we identified 245 and 284 signature genes that are highly expressed in fetal and adult CECs, respectively. Functionally, these genes are enriched in pathways characteristic of CECs, including inorganic anion transmembrane transporter, extracellular matrix structural constituent and cyclin-dependent protein kinase inhibitor activity. Importantly, several of these genes are disease target genes in hereditary corneal dystrophies, consistent with their functional significance in CEC physiology. We also identified stage-specific markers associated with CEC development, such as specific members in the transforming growth factor beta and Wnt signaling pathways only expressed in fetal, but not in adult CECs. Lastly, by the immunohistochemistry of ocular tissues, we demonstrated the unique protein localization for Wnt5a, S100A4, S100A6 and IER3, the four novel markers for fetal and adult CECs. The identification of a new panel of stage-specific markers for CECs would be very useful for characterizing CECs derived from stem cells or ex vivo expansion for cell replacement therapy.

GEO accession number: GSE41616.

INTRODUCTION

Corneal endothelial cells (CECs) are a monolayer of endothelial cells lining the Descemet’s membrane of the cornea. CECs exhibit a typical hexagonal shape and form tight junctions with the neighboring cells. The main function of CECs is to serve as a water pump via Na⁺-K⁺-ATPase, which actively transports bicarbonate ion from the stroma into the anterior chamber. The active bicarbonate ion flux provides the driving force for pumping the corneal fluid out of the stroma, thus keeping the stroma in a constant state of hydration to achieve corneal transparency. Adult CECs have limited capacity for cell proliferation in vivo as the cell cycle is arrested in the G1 phase (1), and CEC deficiency due to genetic diseases, trauma and aging would lead to irreversible corneal edema, opacity and eventual blindness (2).

CEC genetic disorders include Fuchs endothelial corneal dystrophy (FECD), posterior polymorphous corneal dystrophy (PPCD), congenital hereditary endothelial dystrophy type 1 and 2 (CHED1 and 2) and X-linked endothelial corneal dystrophy (3). The genetic defects underlying some of these hereditary diseases have been recently reported. For example, mutations in Collagen type VIII α2 (COL8A2), a major component of the Descemet’s membrane secreted by CECs, are

*To whom correspondence should be addressed. Email: gfan@mednet.ucla.edu (G.F.); deng@jsei.ucla.edu (S.X.D.); and xuezhigang75@gmail.com (Z.X.)

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
associated with either early-onset FEDC or PPDC, whereas mutations in the solute carrier family 4 member A11 (SLC4A11) gene are related to late-onset FEDC and CHED2 (4). Nevertheless, many disease target genes remain to be discovered.

Clinically, besides entire corneal transplantation, surgical procedures such as endothelial keratoplasty or Descemet’s striping endothelial keratoplasty is developed to replace dysfunctional CECs with healthy corneal endothelium (5). However, due to the global shortage of tissue donors and limited transplantation grade cornea tissues, new methods to generate functional CECs need to be established. Recent advancements in stem cell biology have opened new avenues to expand adult or fetal CECs ex vivo (6) or derive CECs from neural crest cells (7), or even with human somatic stem cells such as corneal stroma stem cells, mesenchymal stem cells and bone marrow-derived endothelial progenitor cells (8–10). Theoretically, in vitro differentiation of CECs from pluripotent stem cells such as human embryonic stem cells (hESCs) or induced pluripotent stem cells would be a very feasible approach to provide unlimited source of CECs for cell replacement therapy.

Several protein markers have been used to identify adult CECs, most related to cell adhesion, gap junction, tight junctions and pump function such as N-cadherin, connexin-43 (11), ZO-1 (12), Na⁺–K⁺-ATPase (13) and Occludin (14). However, molecular markers unique to fetal and adult CECs and molecular pathways associated with maturation of CECs are not well studied. We therefore took a comprehensive approach to profile mRNA transcriptomes in both fetal and adult human CECs (hCECs) by RNA-sequencing (RNA-seq) and examine unique molecular markers in fetal and adult CECs. Bioinformatics analysis identified major gene expression changes between fetal and adult CECs in cell metabolism, cell adhesion and the transforming growth factor beta (TGF-β) signaling pathway. By further comparing CEC gene expression profiles with 12 other cell types, we identified 245 and 284 tissue-specific genes in fetal and adult CECs, respectively. In fact, several endothelial dystrophy genes are highly expressed in adult CECs, consistent with their implicated role in CEC physiology. Significantly, we validated a portion of the fetal and adult-specific markers using immunohistochemistry, demonstrating that our data analysis is a valuable resource for identifying protein biomarkers unique to fetal and adult CECs.

Hierarchical clustering revealed that fetal CECs cluster closer to relatively immature cell types, such as hESCs, neural progenitor cells (NPCs) and human umbilical vein endothelial cells (HUVECs; Fig. 1A). In contrast, adult CECs clustered away from these early stage cells and grouped together with other terminally differentiated somatic cell types. By principal component analysis (Fig. 1B), we found that adult CECs are positioned closer to adult brain samples, consistent with the fact that CECs are derived from neural crest cells that share the same cell origin of the neural tube. Overall, identical cell types clustered tightly together, suggesting that each cell type exhibits well-defined transcriptional patterns. In particular, adult and fetal CEC clustered closely together compared with other cell types, indicating that CECs also have a unique transcriptional profile.

To gain insight into the exact molecular differences between adult and fetal CECs from other cell types, we performed statistical analysis (one-way analysis of variance, or ANOVA) and identified genes that were specifically overexpressed in either adult or fetal CECs. Using the criteria of ANOVA Tukey–Kramer multiple comparison corrected P-values <0.01 and an additional cutoff of at least 2-fold greater expression compared with other 12 tissues, we identified a total of 284 (Fig. 2A) and 245 genes (Fig. 2B) that are specifically enriched in adult and fetal CECs, respectively. Gene ontology (GO) analysis revealed that the 284 up-regulated genes in adult CECs mainly participate in inorganic anion transmembrane transporter activity, transcription factor activity and cyclin-dependent protein kinase inhibitor activity (Fig. 3A), reflecting the unique features of mature CECs. Genes unique to fetal CECs are involved in extracellular matrix structural constituent, platelet-derived growth factor binding and growth factor activity (Fig. 3B). These results probably reflect the specific functional traits of fetal CECs for growth factor contribution during fetal CEC development.

We next took a systems approach and asked whether adult or fetal CECs express a unique transcriptional network (or module) of genes. By weighted gene co-expression network analysis (WGCNA), we identified single modules that represent adult (n = 300) and fetal (n = 343) CECs transcriptional networks (Supplementary Material, Figs S1 and S2). When we overlapped this result with our single-gene statistical analysis, we found fairly good concordance with 70% overlap with adult (n = 209) and 44% in fetal (n = 160) CECs (Supplementary Material, Table S1). Taken together, our multi-approach analysis suggests that the unique gene expression profiles in fetal and adult CECs are highly robust.

WGCNA can also lead to a natural measure of gene similarity (or membership) to a particular module. Genes with high membership in a module are generally regarded as hub genes, and tend to play central roles in the network. Interestingly, in adult CECs, we identified SLC4A11 as one of the top-ranking hub genes. SLC4A11 has been linked with hereditary corneal dystrophies (15,16) and knockdown of SLC4A11 in cultured CECs has been shown to induce apoptosis (17). Thus, our network analysis helped to refine and rank the set of key players in the adult or fetal CEC networks. We provided a ranked list of adult and fetal CEC stage-specific hub genes (Supplementary Material, Table S1). Together, our analysis has uncovered a clear transcriptional signature in adult

RESULTS

Tissue-specific gene expression in fetal and adult CECs

Because markers for either fetal or adult CEC are not well studied, we performed RNA-Seq in three adult and two fetal CEC tissues and produced ~179 million uniquely mapped reads. To identify genes that were specific to either fetal or adult CEC, we procured 97 samples from the public domain representing 12 different types of cells and tissues (Supplementary Material, Table S2). For analytical reasons, we only selected RNA-Seq datasets generated by the Illumina system to prevent potential platform biases.
and fetal CECs that are not present in other cell and tissue types examined, including other types of endothelial cells.

**Differential expression pattern between adult and fetal CECs**

We next directly compared adult and fetal CECs and identified 688 and 1763 genes that are expressed by more than 2-fold [false discovery rate (FDR) < 5%] in adult and fetal CECs, respectively. To increase the stringency and identify the most differentially expressed genes between adult and fetal CECs, we used a 5-fold change cutoff, and by this criterion, we identified 518 and 668 genes that are highly expressed in adult and fetal CECs, respectively. Remarkably, functional annotation of the differentially expressed genes suggests distinct functional differences between adult and fetal CECs. For example, GO and kyoto encyclopedia of genes and genomes (KEGG) pathway analysis indicated that many genes involved in metabolic processes are up-regulated in adult CECs (Fig. 4A and Table 1), whereas fetal CECs showed GO terms enriched in metallopeptidase activity, extracellular structure and TGF-β signaling (Fig. 4B and Table 2). Interestingly, calcium ion binding proteins such as S100 family members (S100A4, S100A5 and S100A6) are expressed higher in adult CECs. Collectively, our analysis revealed clear differences between adult and fetal CECs, indicating that CECs acquire transcriptional changes in select pathways during developmental maturation.

**Stage-specific marker genes in adult and fetal CECs are confirmed at the protein level**

Although we have identified many putative molecular markers to distinguish adult and fetal CECs on the RNA level, it is still unclear whether these markers can be confirmed at the protein level. We therefore collected fetal and adult corneal tissues for immunohistochemistry analysis. As expected, immunostaining analysis showed expression of tight junction protein ZO-1 in both fetal (17 weeks in gestation) and adult CECs (>37 years old; Supplementary Material, Fig. S3A and B). Na⁺-K⁺-ATPase, which is one of the most important markers for the pump function in CECs, is highly expressed in adult CECs but not in fetal CECs (Supplementary Material, Fig. S3C and D), consistent with the observations in our RNA-Seq analysis (Supplementary Material, Table S3). This result suggests that pump function is not yet fully established in 17-week-old fetal CECs.

We next examined components of the Wnt signaling pathways since it is known to be an important regulator of ocular tissue development (18,19). Indeed, genes involved in Wnt signaling are highly enriched in fetal CECs (Supplementary Material, Table S1). Our immunocytochemistry experiment showed that Wnt intracellular signaling molecules such...
as axin2 (Supplementary Material, Fig. S3E and F) and beta-catenin (Supplementary Material, Fig. S3G and H) are readily detected in both fetal and adult CECs. However, we found that Wnt5a is only expressed in fetal, but not adult CECs (Fig. 5A and B), suggesting that, at the protein level, the Wnt5a pathway is only functional in fetal CECs.

The S100 protein family consists of several calcium-binding proteins and have been shown to regulate a variety of cellular processes (20). We found that mRNAs in sub-members of the S100A protein family (S100A4, S100A5 and S100A6) are highly expressed in adult, but not fetal CECs (Supplementary Material, Table S1). Our immunohistochemistry confirmed that S100A4 and S100A6 proteins were readily detected in adult CECs (Fig. 5C and E) or fetal cornea epithelium (Supplementary Material, Fig. S4), but not in fetal cornea endothelium (Fig. 5D and F). Therefore, S100A4 and S100A6...
Figure 4. GO analysis of CEC stage-specific signature genes. Bar graphs showing the significance of enrichment terms for a set of higher expressed genes in (A) adult over fetal CECs and (B) fetal over adult CECs. P-values <0.05.

Table 1. Analysis of KEGG pathway via DAVID software for 518 adult stage-specific CECs compared with fetal CECs

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Genes</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis/ Gluconeogenesis</td>
<td>10</td>
<td>ALDOA, GPI, LDHA, ALDOC, PKM2, ENO2, FBP1, PGK1, ALDH3B1, ENO1</td>
<td>1.96E-04</td>
</tr>
<tr>
<td>Arginine and proline metabolism</td>
<td>8</td>
<td>GLS2, SAT1, GLUL, GOT1, ACY1, GATM, NAGS, NOS3</td>
<td>0.002188</td>
</tr>
<tr>
<td>Fructose and mannose metabolism</td>
<td>6</td>
<td>ALDOA, PFKFB4, AKR1B10, PFKFB2, ALDOC, FBP1</td>
<td>0.005918</td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>5</td>
<td>ALDOA, GPI, ALDOC, FBP1, RBKS</td>
<td>0.010124</td>
</tr>
<tr>
<td>p53 signaling pathway</td>
<td>7</td>
<td>STEAP3, TNFRSF10B, CCND3, ZMAT3, CYCS, GADD45G, RRM2B</td>
<td>0.030056</td>
</tr>
<tr>
<td>Nitrogen metabolism</td>
<td>4</td>
<td>GLS2, CTH, GLUL, CA3</td>
<td>0.044035</td>
</tr>
</tbody>
</table>

P-value <0.05.

calcium-binding proteins are two additional novel markers for distinguishing mature adult versus immature fetal CECs.

We also examined cellular localization for proteins that are highly expressed in both adult and fetal CECs. Significantly, we found novel localizations of intracellular proteins such as a stress-induced protein Immediate Early Response 3 (IER3) during CEC maturation. Our immunostaining experiment demonstrated that IER3 can be readily detected in both fetal and adult CECs, but with a stark contrast in subcellular localization. IER3 localizes in the cytoplasm of adult CECs, whereas it is found in the nucleus of fetal CECs. Thus, IER3 is another useful protein marker to characterize the maturation of CECs in development (Fig. 5G and H).

DISCUSSION

CECs are critical for maintaining corneal transparency through regulating stromal hydration. Although previous studies have examined select CEC gene expression via RT-PCR, no information is available concerning the entire transcriptome of either fetal or adult CECs. In this study, we conducted a comprehensive mRNA-Seq analysis of fetal and adult CECs, uncovering both tissue- and stage-specific gene transcripts. These molecular signatures of fetal and adult CECs would be very useful to develop biomarkers to characterize different stages of CECs both in vivo and in vitro.

Here, we have found that genes highly expressed in adult CECs are tightly associated with special CEC functions. For example, the barrier function of CECs is set up by tight junction, which is well correlated with a high level of ZO-1 and N-cadherin mRNAs in adult CECs. Similarly, AQP1, Na\(^+\)-K\(^+-\)ATPase, Na\(^+\)/HCO\(^3\)- and other genes relevant to pump function of mature CECs are detected at a higher level in adult CECs than in fetal CECs. Adult CECs are in a quiescent state with very limited proliferation ability. Consistently, in our analysis, we found that Ki67 and cyclinD2 are dramatically down-regulated in adult CECs compared with fetal CECs.

It has been postulated that both Wnt and TGF-β pathways play a role in CEC development. The Wnt pathways can be separated into two different pathways: canonical frizzled receptor/beta-catemin-dependent T cell factor pathway and non-canonical Ryk receptor/Jun NH2-terminal kinase and calcium/calmodulin-dependent kinase pathways. These two
<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Genes</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECM-receptor interaction</td>
<td>22</td>
<td>COL1A2, COL4A1, COL3A1, COL2A1, COL5A, COL5A1, SDC3, LAMA2, GPS, LAMA4, ITGA5, ITGA8, ITGA7, COL6A3, COL1A2, COL6A2, COL6A1, TNN, COL1A1, THBS2, COL11A1, FN1</td>
<td>1.63E-12</td>
</tr>
<tr>
<td>Axon guidance</td>
<td>22</td>
<td>PLXNC1, PLXNA1, GNA11, EFNB1, EFNB2, EphB3, CXC12L, EphA3, SLIT3, EphB2, SEMA5B, EphB6, SEMA6B, EphA6, CXCR4, ROBO1, FYN, SEMA3F, SEMA3D, ROBO2, SEMA3A, SRGAP1</td>
<td>8.34E-09</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>26</td>
<td>COL3A1, COL2A1, BCL2, COL6A3, COL6A2, COL6A1, TNN, PIK3R3, THBS2, COL11A1, FN, COL4A2, COL4A1, IGFL1, FLNC, COL5A2, COL5A1, LAMA4, ITGA5, FYN, CCND2, ITGA8, ITGA7, COL1A2, COL1A1</td>
<td>8.22E-08</td>
</tr>
<tr>
<td>Arrhythmogenic right ventricular cardiomyopathy</td>
<td>10</td>
<td>LAMA2, SLCA8A1, CACNA2D1, PKP2, ITGA5, ITGA8, ITGA7, LEF1, CACNG4, SGCd</td>
<td>0.001988</td>
</tr>
<tr>
<td>Hypertrophic cardiomyopathy</td>
<td>10</td>
<td>LAMA2, SLCA8A1, CACNA2D1, ITGA5, ITGA8, ITGA7, IGFl1, CACNG4, SGCd, TGFβ2</td>
<td>0.004299</td>
</tr>
<tr>
<td>Pathways in cancer</td>
<td>23</td>
<td>FGFR1, RET, AR, COL4A2, BMP2, COL4A1, LEF1, KITLG, IGFl1, CDK6, FGFR2, GLI2, MMP2, FZD7, TGFβ2, LAMA2, SMO, FZD10, CBLB, LAMA4, BCL2, PIK3R3, FN1</td>
<td>0.005296</td>
</tr>
<tr>
<td>Dilated cardiomyopathy</td>
<td>10</td>
<td>LAMA2, SLCA8A1, CACNA2D1, ITGA5, ITGA8, ITGA7, IGFl1, CACNG4, SGCd, TGFβ2</td>
<td>0.007243</td>
</tr>
<tr>
<td>Cell adhesion molecules</td>
<td>12</td>
<td>NCAM2, CADM3, CLDN19, CD34, ITGA8, CD276, NLGN1, NLGN3, CDH3, JAM2, SDC3, CLDN15</td>
<td>0.010309</td>
</tr>
<tr>
<td>Small cell lung cancer</td>
<td>8</td>
<td>LAMA2, LAMA4, COL4A2, COL4A1, BCL2, CDK6, PIK3R3, FN1</td>
<td>0.037463</td>
</tr>
<tr>
<td>TGF-beta signaling pathway</td>
<td>8</td>
<td>BMP2, LTBP1, ACVR1L1, LEFT2Y2, DCN, THBS2, CHRD, TGFβ2</td>
<td>0.044046</td>
</tr>
<tr>
<td>Gap junction</td>
<td>8</td>
<td>TUBB2B, GNA11, TUBB6, LPAR1, PLCB1, ITPR3, TUBA1A, TUBB4</td>
<td>0.048826</td>
</tr>
</tbody>
</table>

P-value < 0.05.

Pathways are targeted by different Wnt molecules such as Wnt1, Wnt3a and Wnt8 for the canonical pathway, whereas Wnt5a, Wnt4 and Wnt11 for the non-canonical pathway (21). Of note, we found that 27 genes in the Wnt pathway are highly expressed in fetal CECs (Supplementary Material, Tables S1 and S4), particularly for the non-canonical Wnt pathway genes such as Wnt5a and Wnt4, raising the possibility that the non-canonical Wnt pathway may be important for CEC development. Similarly, eight TGF-β signaling pathway members (TUBB2B, GNA11, TUBB6, LPAR1, PLCB1, ITPR3, TUBA1A and TUBB4) are detected at a much higher level in fetal CECs than in adult CECs. TGF-β family genes are involved in regulating cell differentiation, proliferation and extracellular matrix production in ocular tissue (22). In fact, TGF-β2 null mice exhibit an absence of endothelium and fusion of the lens with the cornea (23). In coupling with high levels of TGF-β signaling molecules, we also observed enrichment of genes in the extracellular matrix structural constituent in fetal CECs. Overall, our stage-specific gene expression analysis supports the notion that both Wnt and TGF-β pathways are important for the differentiation and maturation of fetal CECs into adult CECs.

Recently, derivation of CEC-like cells in vitro from progenitor cells or somatic stem cells has been investigated for transplantation treatment of CEC deficiency. Hatou et al. (8) isolated cornea-derived precursors from the adult corneal stroma to make tissue-engineered corneal endothelium. Joyce et al. (9) investigate the potential of umbilical cord blood mesenchymal stem cells to become HCECs. Recently, Ju et al. (7) tested the feasibility of differentiating neural crest cells into functional corneal endothelial cells. In all these studies, the markers used for characterizing CEC-like cells were limited to ZO-1, Na⁺-K⁺-ATPase and N-cadherin. Our demonstration of four novel protein markers including S100A4, S100A6, Wnt5a and IER3 in this study will certainly aid the characterization of CEC-like cells from the above studies. S004A4 is one of S100 family genes encoding calcium-binding proteins involved in cell morphogenesis in multiple cell lineages (24). S100A6 was previously shown in the plasma membrane of corneal and limbal epithelial cells (25). It has been debated whether S100 expression is specific in human corneal endothelium (26,27). In the present study, we found that S100A4 and S100A6 are detectable in adult CECs as well as in adult and fetal corneal epithelium, but not in fetal CECs. Thus, S100A4 or S100A6 should be considered as a mature marker for adult CECs. IER3 is a stress-induced protein (28) encoded by radiation-inducible immediate-early gene IEX-1, previously detected as a potential target gene in human cancer, inflammatory diseases or hypertension (29,30). In keratinocytes, it is observed that IER3 can translocate from nucleus to cytoplasm in response to 1alpha,25-dihydroxyvitamin D3 for reduced cell growth (31). Consistently, our observed IER3 translocation from the nucleus in fetal CECs to the cytoplasm in adult CECs may also be associated with reduced cell proliferation and growth in adult CECs. Taken together, our newly identified CEC protein markers as well as stage-specific CEC mRNAs would be very useful to characterize and quality-control a variety of CEC-like cells derived from either ex vivo expansion of corneal progenitors or in vitro differentiation of somatic and pluripotent stem cells.

Lastly, corneal endothelial dystrophies are a set of hereditary corneal diseases characterized by slowly progressive edema of the cornea in different modes, each associated with different target genes. We noted that several of our newly identified CEC-tissue-specific genes are also corneal disease target genes. For example, COL8A2 and SLCA411, two major PPCD-related genes (32), are expressed at an extremely high level in both fetal and adult CECs. Similarly, CDKN1A exhibits a much higher expression level in adult CECs than fetal CECs. CDKN1A is an important marker of cellular senescence (33) and mediates part of the cellular stress response in the pathogenesis of both early and late-onset FECD (34). Our current study identifies over two hundred
genes that are highly expressed in adult and fetal CECs when compared with 12 other tissue types (Supplementary Material, Table S2). Because these CEC signature genes are enriched for the genes critical for CEC functions, we speculate that some of these highly expressed signature genes in adult and fetal CECs could be good candidates of disease target genes underlying many other hereditary CEC dystrophies.

MATERIALS AND METHODS

Adult and fetal CEC samples

Adult cornea samples (age 31, 56 and 64 years old) were obtained from normal subjects who donated to the eye bank in UCLA Jules Stein Eye Institute. Fetal CEC samples were dissected from fetal ocular samples (16–18 weeks of gestation) provided by UCLA Department of Pathology & Laboratory as approved by UCLA IRB. The whole corneal endothelial layers were carefully peeled off under a dissecting microscope and were immediately lysed in reagents for RNA purification.

RNA isolation

RNA were extracted from those adult and fetal tissues using the RNasy Micro Kit (Qiagen, USA). RNA concentration was detected by nanodrop spectrophotometer.

Library construction and high-throughput sequencing

RNA-Seq library construction followed the protocol described in Illumina TruSeq™ RNA Sample Preparation Guide. Library construction was started with 100 ng of total RNA, via poly-T oligo-attached magnetic beads to purify the poly-A containing mRNA molecules. RNA fragments were reverse-transcribed into first cDNA strand, followed by
synthesis second cDNA strand. After end repaired with a single ‘A’ base, cDNAs were ligated with different adapters for PCR amplification. After being amplified for 15 cycles, the concentration of the product was tested by the Qubit fluorometer (Invitrogen, USA). We subjected 10 nmol of each sample to sequencing according to the manufacturer’s instruction with Illumin Hi-seq 2000. Reads were mapped to the hg19 genome using burrows-wheeler alignment (35).

Statistical and bioinformatics analysis

For all datasets analyzed, we first transformed mapped transcript read counts to the reads per kilobase per million mapped reads (RPKM) metric, and genes with an average RPKM < 1 were filtered out, followed by quantile normalization. We performed Student’s t-test with Storey’s multiple testing correction to produce the FDR. Genes that had greater than a mean fold-change of 2 and FDR < 5% were considered differentially expressed. All statistical analysis was performed using built-in Matlab functions. GO and KEGG pathway analyses were performed via David bioinformatics on-line software.

Tissue-specific expression

We curated multiple samples from the public domain, restricting ourselves to datasets generated by the Illumina Solexa platform to reduce platform biases. We carried out one-way ANOVA, and then selected for genes that were specifically overexpressed in either adult or fetal CEC samples. Two cutoffs were used: (1) Tukey–Kramer multiple comparison corrected P-values <0.01, which is a relatively conservative method for one-way ANOVA with different sample sizes and (2) mean 2-fold expression over all tissue types. These analyses were performed in Matlab using built-in functions (anova1, and multcomp).

Weighted gene co-expression network analysis

We constructed a signed weighted correlation network as previously described (36). Briefly, a matrix of pairwise correlations was generated between all pairs of genes across all samples. This matrix was raised to the power 12 to produce the adjacency matrix, which is used to calculate the Topological Overlap (TO), a robust and biologically meaningful measure of network interconnectedness. Genes with highly similar co-expression relationships were grouped together by performing average linkage hierarchical clustering on the TO. Module membership was calculated for each gene i in module q, as MMiq = cor(xi, Eq), where xi is the expression profile of gene i and Eq is the eigengene of module q. Genes with high module memberships are generally regarded as hub genes. WGCNA is a publicaly available R package that has been described in detail in the reference (36).

Immunostaining

Adult and fetal cornea tissues were fixed with 4% PFA for 30 min at room temperature, and cryo-protected in 30% sucrose/PBS overnight. After embedded in OCT, frozen section were cut using a microtome-cryostat (Leica, USA) in 10 μm. The sections were washed three times in phosphate-buffered saline (PBS) and then permeabilized in 0.4% Triton-X 100 for 1 h. Blocking was carried out using 4.5% normal donkey serum, then sections were incubated overnight at 4°C in primary antibodies diluted in blocking solution. The primary antibodies used in the experiment were anti-Wnt5a (Santa Cruz), S100A4 (Abnova), S100A6 (Abnova), IER3 (Abgent), ZO-1 (Invitrogen), Na+–K+–ATPase (Millipore), Axin2 (abcam) and Beta-catenin (BD). After being washed three times in PBS, sections were incubated in fluorochrome-conjugated secondary antibodies for 1 h at room temperature, cell nuclei staining was done by Hoechst, and the slides were mounted with 5% n-propylgallate (Sigma) in 50% glycerol. Immunostaining results were taken using a 40× objective under fluorescent light with a confocal microscope (Leica TCS-SP; Leica Microsystems).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We acknowledge the assistance of UCLA Broad Stem Cell Research Center Sequencing Core for high-throughput sequencing.

Conflict of Interest statement: None declared.

FUNDING

This work was supported by NIH grant GM081621, International Science & Technology Cooperation Program of China #2011DFB30010, Ministry of Science and Technology 973 Basic Research program in China (2011CB965102, 2011CB966204 and 2012CB966303), National Natural Science Foundation of China (81271258). Y.C. is partly supported by Chinese Scholarship Council (CSC) funding (No. 2011626117). K.H. is supported by the California Institute of Regenerative Medicine Training Grant (TG2-01169).

REFERENCES


Chapter 6
Identification of miRNA Signatures during the Differentiation of hESCs into Retinal Pigment Epithelial Cells

Ganlu Hu1,2*, Kevin Huang1, Juehua Yu1, Saiselakshmi Pillai4, Jun Kong1, He Xu1, Zhenshan Liu2, Kunshan Zhang5, Jun Xu6, Yuping Luo5, Siguang Li5, Yi E. Sun1,2, Linda E. Iverson6, Zhiqiang Xue1,2, Guoping Fan1,2,3

1Translational Center for Stem Cell Research, Tongji Hospital, Tongji University School of Medicine, Shanghai, China, 2Stem Cell Research Center, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China, 3Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America, 4Stem Cell Biology, Beckman Research Institute of City of Hope, Duarte, California, United States of America

Abstract
Retinal pigmented epithelium (RPE) cells can be obtained through in vitro differentiation of both embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs). We have previously identified 87 signature genes relevant to RPE cell differentiation and function through transcriptome analysis of both human ESC- and iPSC-derived RPE as well as normal fetal RPE. Here, we profile miRNA expression through small RNA-seq in human ESCs and their RPE derivatives. Much like conclusions drawn from our previous transcriptome analysis, we find that the overall miRNA landscape in RPE is distinct from ESCs and other differentiated somatic tissues. We also profile miRNA expression during intermediate stages of RPE differentiation and identified unique subsets of miRNAs that are gradually up- or down-regulated, suggesting that dynamic regulation of these miRNAs is associated with the RPE differentiation process. Indeed, the down-regulation of a subset of miRNAs during RPE differentiation is associated with up-regulation of RPE-specific genes, such as RPE65, which is exclusively expressed in RPE. We conclude that miRNA signatures can be used to classify different degrees of in vitro differentiation of RPE from human pluripotent stem cells. We suggest that, RPE-specific miRNAs likely contribute to the functional maturation of RPE in vitro, similar to the regulation of RPE-specific mRNA expression.


Editor: Martin Pera, University of Southern California, United States of America

Received: January 23, 2012; Accepted: April 16, 2012; Published: July 27, 2012

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: This work was supported by 2011CB966102, 2011CB966204, and 2012CB966300 grants from the Ministry of Science and Technology in China, 2011T0700010 from the International Science & Technology Cooperation Program of China, National Institutes of Health Grants P01GM081521 and R21NS097924, and RC1-0911 and TR1-272 from the California Institute for Regenerative Medicine. Kevin Huang and Juehua Yu are supported by a California Institute of Regenerative Medicine Training Grant (T02-01-169). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: GHu@mednet.ucla.edu (GH); Xuexihuang785@gmail.com (ZX)

These authors contributed equally to this work.

Introduction
The retinal pigment epithelium (RPE) is a polarised monolayer of pigmented cells between the neural retina and choroid. RPE plays many crucial roles in the eye including blood/retina barrier formation, nutrient transport between blood to photoreceptors, water conveyance from subretinal space to the blood, light absorption, growth factor secretion, and phagocytosis of the outer segment of photoreceptors [1]. For these reasons, dysfunction or death of RPE can lead to photoreceptor degeneration and eventual blindness in diseases such as Stargardt disease and age-related macular degeneration (AMD) [2,3,4]. Although there are some treatments that partially alleviate symptoms of the wet form of AMD, there is no cure for AMD, particularly the more prevalent dry form of AMD [5].

Recent advances in stem cell biology have motivated new strategies to develop pluripotent stem-cell derived RPE for cell replacement therapy of AMD [6]. We and other labs have achieved the ability to generate functional RPE through in vitro differentiation of both human embryonic stem cells (hESCs) and induced pluripotent stem cells (iPSCs) [7,8,9,10,11,12]. However, the process can take between three to six months to achieve functional stem cell-derived RPE. Other approaches for promoting RPE differentiation through defined factors or small molecules such as ricanidomide [7,11,13] However, our understanding of the molecular changes associated with RPE differentiation are still limited.

A recent microarray study comparing gene expression profiles of human fetal and adult RPE with somatic tissues identified 134 signature genes that are unique to RPE [14]. Following this study, we profiled miRNA expression during in vitro differentiation of hESCs/iPSCs into RPE, By comparing the expression patterns of 134 RPE signature genes between hESCs/iPSCs versus stem cell-derived RPE, we refined the list of signature genes to a set of 87 RPE specific genes that can be used to monitor RPE...
differentiation and distinguishes stem cell-derived RPE from other cell types.

In this current study, we followed the microRNA expression patterns during the course of RPE differentiation at four separate time points to survey molecular changes associated with RPE differentiation. MicroRNAs are a class of 18–26nt length small RNAs which can attenuate gene expression by inhibiting translation in the cytoplasm or promoting microRNA degradation in the nucleus [15]. There is growing evidence suggesting microRNAs play a role in RPE differentiation, survival, and function. For example, Dicer conditional knockout mice have decreased mature microRNA expression in retinas, leading to retinal degeneration and severely impaired visual function [16]. Therefore, elucidating the functional role of microRNAs during RPE differentiation will give us important insights into the crucial microRNAs involved in promoting RPE differentiation and maturation.

Results

RNA-seq analysis of microRNA profiles in human stem cells, RPE, and other types of somatic cells

We have developed a protocol to derive functional RPE cells from human pluripotent stem cells through a time window of three to six months or more [9]. Previously, we have demonstrated that stem cell-derived RPE are functionally equivalent to fetal RPEs and share global microRNA expression profiles with cultured primary human RPE [9]. To profile microRNA expression patterns, we constructed small RNA libraries from 1) human ESCs, 2) 15-day partially differentiated hESCs, 3) early pigmented clusters (PC) that appear around 30 days, and 4) RPE cells in culture for over 3 months [9,10]. Using high-throughput, we obtained a high-resolution microRNA profile in these cells. We also incorporated microRNA sequencing datasets of various human somatic tissues including heart, breast, skin and melanocytes procured from the Gene Expression Omnibus (GEO) database (Table S1).

Unbiased hierarchical sample classification analysis and principal component analysis (PCA) both revealed distinct microRNA expression patterns in ESCs, partially differentiated cells, fully differentiated RPE cells, and other somatic tissues (Fig 1B, 1C). Interestingly, hierarchical clustering showed that HSF1-RPE and fetal RPE cluster closer to immature pigment clusters, indicating this particular batch of HSF1-RPE may not be fully mature compared to H9-RPE and hiPS2-RPE. In addition, our data indicated that partially differentiated hESCs and immature RPE shared many similar microRNA characteristics, consistent with the notion that microRNA signatures can be used to gauge the state of RPE differentiation. 3D-profiles of the first three principal components further confirmed our hypothesis that HSF1-RPE is immature compared to H9-RPE and hiPS2-RPE, which cluster away from ESC and other partially differentiated cells and closer to fully differentiated somatic tissues. Importantly, RPE clustered away from melanoblast and melanocyte cells, another type of pigment cell also derived from neural crest. This is consistent with our previous study demonstrating distinct transcriptional differences between RPE and melanocytes [9], and suggests pigmentation between these two cell types may not be regulated in the same fashion or that the overall cell-type specific microRNA profile dominates the majority of differences between the two pigmented cell types. Moreover, we found RPE clustered away from other epithelial cell types, including breast and skin tissues and the human mammary epithelial cell line (HMEC), indicating RPE express a unique microRNA profile different from other epithelial cells. Collectively, our results revealed a distinct set of microRNAs for RPE compared to ESCs cells and other somatic tissues.

Figure 1. ES-derived RPE have distinctive global microRNA signatures. A) Morphological features of four stages during human pluripotent stem cells differentiate into RPE cells: a) human ES cells, b) partially differentiated ES cells after 4 days withdrawal bFGF, c) pigment cluster appeared after about 30 days differentiation; d) RPE monolayer. Scare bar: (a-c) 100 μm; (d) 50 μm. B) C) RNA profiles of different cell types were clustered using either unbiased hierarchical clustering or principal component analysis (PCA). B) Pearson correlation clustering between all cell lines. C) 3D-biplots of the first three principal components. Various somatic tissues were downloaded from GEO database.

doi:10.1371/journal.pone.0037224.g001
Figure A: Graphs showing expression levels of different miRNAs across various cell types.

Figure B: Heatmap illustrating the expression pattern of miRNAs in different cell stages.

Figure C: Bar chart displaying fold mRNA expression over different cell types and conditions.
Figure 2. Select miRNAs are dynamically expressed during differentiation of H9 ES cell line. Representative miRNA expressions that are (A) down-regulated or (B) up-regulated during the course of differentiating H9 ESCs. All values were normalized to ES state, then log2 transformed. (C) RT-PCR analysis of different RPE differentiation stage showing miR-23a, miR-125a, miR-125b, miR-302d and let-7g expression relative to H9 ESC stage. The relative expression levels for each miRNA were normalized by hsa-U6 control. PD = partially differentiated, PC = pigmented clusters, ES = embryonic stem cells.

doi:10.1371/journal.pone.0037224.g002

miRNA dynamics during RPE differentiation

To examine miRNA dynamics during RPE differentiation, we compared miRNA profiles during the course of RPE differentiation at four distinct stages: ESC cells (stage 1), 13 days partially differentiated ES cells (stage 2), early pigmented clusters after 30-day differentiation (stage 3), and RPE (stage 4). We used one-way ANOVA test to determine whether miRNA expression within our group was significantly differentially expressed compared to the other three groups. In total, we found 21 and 49 miRNAs that were significantly up- and down-regulated in RPE compared to ES cells (p<0.05) (Figure 2, Table 1, 2). In particular, we found two groups of miRNAs that were dynamically regulated. One group of miRNAs was significantly and progressively up-regulated during differentiation from hESCs/iPSCs to 15 days differentiation, pigmented clusters, and all the way to fully differentiated RPE. Similarly, we found that another set of miRNAs was gradually down-regulated during the course of RPE differentiation (Figure 2A). As expected, the miRNAs we identified were mostly involved in differentiation. For example, miRNAs that promote differentiation such as let-7 family were up-regulated whereas miRNAs that promote pluripotency such as the miR-300 cluster are down-regulated. We were able to validate the expression of these miRNAs and others using Taqman qPCR, confirming our small RNA-Seq quantitation (Figure 2C). Thus, our data indicated that the majority of miRNA changes during RPE differentiation are involved in silencing of pluripotency-related miRNAs and activation of differentiation miRNAs common to most somatic cells. Interestingly, we also found a group of miRNAs that were specifically enriched at stage 2 and 3 of differentiation, but depleted in both stage 1 and 4 (Figure 2B). KEGG pathway analysis of the predicted targets of these miRNA was most involved in endocytosis, active transport, lysosome and other RPE related function (Figure S1). These miRNA may be involved in early phase of hESC/RPE differentiation.

To examine the predicted targets of down-regulated miRNAs, we used two commonly used miRNA target prediction software TargetScan and miRanda. We compiled a list of non-redundant predicted targets, combined outputs from both softwares, and found several thousand putative targets. Gene Ontology (GO) analysis revealed predicted targets were enriched for biological functions involved in cell adhesion, cytoskeleton and other RPE related processes.

Table 1. Downregulated miRNAs in RPEs during ESC differentiation.

<table>
<thead>
<tr>
<th>miRNA</th>
<th>H9 ESC</th>
<th>H9 RPE</th>
<th>Log2fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-106b</td>
<td>2779</td>
<td>147</td>
<td>-4.2</td>
</tr>
<tr>
<td>hsa-mir-130a</td>
<td>1032</td>
<td>61</td>
<td>-4.1</td>
</tr>
<tr>
<td>hsa-mir-130b</td>
<td>521</td>
<td>19</td>
<td>-4.8</td>
</tr>
<tr>
<td>hsa-mir-1323</td>
<td>106</td>
<td>1</td>
<td>-5.4</td>
</tr>
<tr>
<td>hsa-mir-15a</td>
<td>166</td>
<td>30</td>
<td>-2.5</td>
</tr>
<tr>
<td>hsa-mir-15b</td>
<td>277</td>
<td>15</td>
<td>-4.2</td>
</tr>
<tr>
<td>hsa-mir-16-2</td>
<td>549</td>
<td>240</td>
<td>-1.2</td>
</tr>
<tr>
<td>hsa-mir-17</td>
<td>5685</td>
<td>694</td>
<td>-3.0</td>
</tr>
<tr>
<td>hsa-mir-18a</td>
<td>1387</td>
<td>44</td>
<td>-5.0</td>
</tr>
<tr>
<td>hsa-mir-25</td>
<td>440</td>
<td>84</td>
<td>-2.4</td>
</tr>
<tr>
<td>hsa-mir-301a</td>
<td>88</td>
<td>6</td>
<td>-3.8</td>
</tr>
<tr>
<td>hsa-mir-302c</td>
<td>9948</td>
<td>1</td>
<td>-13.3</td>
</tr>
<tr>
<td>hsa-mir-302d</td>
<td>6481</td>
<td>1</td>
<td>-12.4</td>
</tr>
<tr>
<td>hsa-mir-32</td>
<td>236</td>
<td>11</td>
<td>-4.5</td>
</tr>
<tr>
<td>hsa-mir-22a</td>
<td>62</td>
<td>3</td>
<td>-4.2</td>
</tr>
<tr>
<td>hsa-mir-363</td>
<td>3845</td>
<td>21</td>
<td>-7.5</td>
</tr>
<tr>
<td>hsa-mir-373</td>
<td>450</td>
<td>1</td>
<td>-8.8</td>
</tr>
<tr>
<td>hsa-mir-374a</td>
<td>2329</td>
<td>66</td>
<td>-5.1</td>
</tr>
<tr>
<td>hsa-mir-512-1</td>
<td>399</td>
<td>1</td>
<td>-8.6</td>
</tr>
<tr>
<td>hsa-mir-516b-2</td>
<td>130</td>
<td>1</td>
<td>-7.0</td>
</tr>
<tr>
<td>hsa-mir-548f-1</td>
<td>361</td>
<td>1</td>
<td>-8.5</td>
</tr>
<tr>
<td>hsa-mir-598</td>
<td>555</td>
<td>2</td>
<td>-7.5</td>
</tr>
<tr>
<td>hsa-mir-93</td>
<td>2340</td>
<td>377</td>
<td>-2.6</td>
</tr>
<tr>
<td>hsa-mir-86</td>
<td>174</td>
<td>2</td>
<td>-6.6</td>
</tr>
</tbody>
</table>

Numbers are normalized miRNA counts in each library.
doi:10.1371/journal.pone.0037224.t001

Table 2. Upregulated miRNAs in RPE during ESC differentiation.

<table>
<thead>
<tr>
<th>miRNA</th>
<th>H9 ESC</th>
<th>H9 RPE</th>
<th>Log2fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa-mir-7a-2</td>
<td>305</td>
<td>9107</td>
<td>4.9</td>
</tr>
<tr>
<td>hsa-mir-7c</td>
<td>20</td>
<td>1288</td>
<td>5.0</td>
</tr>
<tr>
<td>hsa-mir-7f-1</td>
<td>115</td>
<td>11339</td>
<td>6.6</td>
</tr>
<tr>
<td>hsa-mir-7f-2</td>
<td>118</td>
<td>11685</td>
<td>6.6</td>
</tr>
<tr>
<td>hsa-mir-7g</td>
<td>11</td>
<td>11615</td>
<td>10.0</td>
</tr>
<tr>
<td>hsa-mir-7i</td>
<td>7</td>
<td>9875</td>
<td>10.5</td>
</tr>
<tr>
<td>hsa-mir-100</td>
<td>13</td>
<td>6309</td>
<td>8.9</td>
</tr>
<tr>
<td>hsa-mir-125a</td>
<td>37</td>
<td>19072</td>
<td>9.0</td>
</tr>
<tr>
<td>hsa-mir-125b-1</td>
<td>19</td>
<td>39452</td>
<td>11.0</td>
</tr>
<tr>
<td>hsa-mir-125b-2</td>
<td>18</td>
<td>38627</td>
<td>11.1</td>
</tr>
<tr>
<td>hsa-mir-139</td>
<td>5</td>
<td>971</td>
<td>6.2</td>
</tr>
<tr>
<td>hsa-mir-151b</td>
<td>15</td>
<td>57</td>
<td>2.5</td>
</tr>
<tr>
<td>hsa-mir-152</td>
<td>13</td>
<td>2087</td>
<td>7.3</td>
</tr>
<tr>
<td>hsa-mir-181d</td>
<td>1</td>
<td>122</td>
<td>6.9</td>
</tr>
<tr>
<td>hsa-mir-214</td>
<td>2</td>
<td>602</td>
<td>8.2</td>
</tr>
<tr>
<td>hsa-mir-22</td>
<td>25</td>
<td>2563</td>
<td>6.7</td>
</tr>
<tr>
<td>hsa-mir-222</td>
<td>167</td>
<td>1540</td>
<td>3.3</td>
</tr>
<tr>
<td>hsa-mir-23a</td>
<td>32</td>
<td>3188</td>
<td>5.9</td>
</tr>
<tr>
<td>hsa-mir-26a-1</td>
<td>636</td>
<td>6842</td>
<td>3.4</td>
</tr>
<tr>
<td>hsa-mir-32a-2</td>
<td>639</td>
<td>6984</td>
<td>3.4</td>
</tr>
<tr>
<td>hsa-mir-3120</td>
<td>2</td>
<td>602</td>
<td>8.2</td>
</tr>
<tr>
<td>hsa-mir-4521</td>
<td>16</td>
<td>156</td>
<td>3.3</td>
</tr>
<tr>
<td>hsa-mir-98</td>
<td>3</td>
<td>930</td>
<td>8.3</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0037224.t002
Table 3. RPE signature genes associated with downregulated miRNAs.

<table>
<thead>
<tr>
<th>ADAM9</th>
<th>MIRBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANKRD12</td>
<td>NAV1</td>
</tr>
<tr>
<td>CDH1</td>
<td>NEDD4L</td>
</tr>
<tr>
<td>CDH2</td>
<td>PTPRA</td>
</tr>
<tr>
<td>ECH2</td>
<td>PANK2</td>
</tr>
<tr>
<td>CER1</td>
<td>IAB3B</td>
</tr>
<tr>
<td>DDX1</td>
<td>REH(183)</td>
</tr>
<tr>
<td>DDX1L1</td>
<td>REH(183)</td>
</tr>
<tr>
<td>ENPP2</td>
<td>IBAGD</td>
</tr>
<tr>
<td>FBP2</td>
<td>SDC2</td>
</tr>
<tr>
<td>G3S1</td>
<td>SGK3</td>
</tr>
<tr>
<td>IFG4</td>
<td>SMAD6</td>
</tr>
<tr>
<td>ITM2B</td>
<td>SORBS2</td>
</tr>
<tr>
<td>MIR21L1</td>
<td>TIMP3</td>
</tr>
<tr>
<td>MIR212</td>
<td>TPP1A</td>
</tr>
<tr>
<td>MET</td>
<td>TIP3</td>
</tr>
<tr>
<td>MIFAP3L</td>
<td>VEGA</td>
</tr>
</tbody>
</table>

Conservation of regulatory network between human mouse RPE miRNA-signature genes

Because many core biological functions are conserved between human and mouse, we cross-referenced our data with a previously published mouse RPE miRNA profiling dataset [20] and used NAVIGATOR [21] to analyze the complex relationships between miRNA and RPE core genes. We identified 9 conserved miRNAs that appeared in top 100 most abundant miRNAs in both the human and mouse RPE datasets. After comparing their predicted targets to 87 RPE signature genes, we obtained a network map between homologous miRNAs and human RPE signature genes (Figure 4). In this network, one miRNA can target several genes and one gene can be targeted by several miRNAs. We found 8 RPE signature genes were targeted by both human and mouse miRNAs. 5 genes were exclusively targeted by human miRNA and 1 gene is exclusively targeted by mouse miRNA. GO analysis of shared genes has the same function both in human and mouse. Notably, 7 of these conserved miRNA are to be specifically expressed in RPE compared to other tissues, indicating expression of these key miRNAs is critical for defining RPE cells.

Discussion

The RPE is a highly specialized epithelium that is crucial for maintaining photoreceptor function and visual cycle [1]. Previous studies have shown a critical role for miRNAs in eye development, function, and survival in zebrafish-deficient mice show multiple defects, indicating an important role for miRNAs in many tissues including the eye [24,25,26]. Previous miRNA profiles in RPE have primarily focused on mouse RPE or the human cell line ARE19 [25-27], however, miRNA profiles in human RPE are still limited. MiRNAs are proposed to act as fine tuners of gene expression either through inhibiting translation in the cytoplasm, or promoting mRNA degradation in the nucleus [27,28]. More recently, miRNAs have been shown to primarily repress genes through transcriptional control [29]. Therefore, we used different software to predict the miRNA target of differentially expressed miRNA. Many of the down-regulated miRNAs potentially target a generous portion of RPE signature genes. However, a recently published study comparing various miRNA prediction softwares show very low concordance between three major prediction algorithms (miRanda, TargetScan, PicTar), indicating that the standard for predicting miRNA target genes is still not robust [30]. To the best of our abilities and available tools, we employed these prediction algorithms to provide putative targets of key miRNAs. Through profiling miRNA expression at four defined time points during RPE differentiation, our study demonstrated that each stage of RPE differentiation has a unique subset of miRNAs that are significantly differentially expressed compared to all other stages. We identified miRNAs that were exclusively enriched in ESCs, intermediate stages of spontaneously differentiated ESCs, and fully derived RPE. In addition, we were able to show that a portion of the miRNA become gradually increased or decreased, or transiently increased during the differentiation process, suggesting the expression level of particular miRNAs may be used as an indicator for RPE maturation. Thus, our study indicated that the degree of RPE differentiation can be gauged by profiling the specific miRNA expression patterns during RPE differentiation as shown in Figure 2. Furthermore, by incorporating miRNA profiles
from various somatic tissues, we were able to find a portion of miRNA that are specifically expressed in human RPE cells. Functional annotation of the predicted targets using DAVID revealed RPE-specific miRNAs are primarily associated in regulating the epithelial barrier and TGF-β pathway. TGF-β signaling pathway is essential for epithelium-mesenchymal transition in RPE cells, indicating suppression of the TGF-β pathway contributes to maintaining the epithelium in RPE [31,32]. Future study shall validate the function of these miRNAs in RPE differentiation through gain or loss of function experiments. We also find a portion of human RPE-specific miRNAs also share enrichment in mouse RPE, indicating a potentially conserved functional role for these miRNAs. We suggest that the human RPE-specific miRNA signature may serve as molecular makers for characterizing functional RPE.

Age-related macular degeneration (AMD) is characterized by malfunction and degeneration of RPE. Recently, Kameko et al. [33] discovered a miRNA-independent cell survival function for DICER1 in the context of AMD pathology, demonstrating upregulation of Alu elements in the absence of DICER1 promoted RPE cell death. However, this does not exclude the possibility for miRNAs in maintaining other features of RPE. Indeed, other studies have shown that miRNA may play important roles in AMD pathogenesis. Lin et al. discovered that miR-23a was down
regulated in AMD eyes while upregulation of miR-25a can protect RPE cell from oxidative damage in ARPE-19 cells [34]. In our dataset, miR-25a was enriched in RPE cells, and its expression was significant increased during development, supporting the hypothesis that miR-25a expression is necessary for maintaining healthy RPE. Overall, our data uncovered a unique set of miRNAs that are expressed in RPE, suggesting a small number of key miRNAs may contribute to promoting RPE survival and function.

Materials and Methods

Differentiation of hESC/hiPSC-RPE

The human embryonic stem cell line H9, HSF1 as well as induced pluripotent stem cell line hPS2 and mouse embryonic feeder cells were obtained from the UCLA Stem Cell Core [9]. Human embryonic stem cell line hES1 was a generous gift from Dr. James Thomson [18]. Pluripotent stem cells (hESC and iPSCs) were plated onto gamma-rays irradiation mouse embryonic feeder cells with DMEM/F12 culture medium containing 20% Knock- Out Serum Replacement, 0.1 mM nonessential amino acids, 0.1 mM b-mercaptoethanol and 100 ng/ml zebrafish basic fibroblast growth factor (zhbFGF) on a 6-well plate. Briefly, cells were cultured at 37°C in 5% CO2 for 6-10 days after which zhbFGF was omitted to facilitate spontaneous cell differentiation.

Pigmented colonies were observed within 4 weeks and allowed to expand for a few weeks, with media changes every 2-3 days. Pigmented cells were enriched by manual dissection using insulin needle followed by seeded on growth factor reduced Matrigel/BioFluorescent, diluted 1:30 coated plate and transwell membranes.

RPE medium were changed to support pigment cluster expansion (containing xMEM, 1 x N2 supplement(Gibco), 1 x Non-essential amino acid solution, 250 mg/ml taurine, 15 mg/ml Triliodo thyronine (Sigma-Aldrich, Gillingham, UK), 1 L L-glutamine (Invitrogen, Paisley, UK), 1 L Penicillin-streptomycin and 10% Hydrolone heat-inactivated foetal bovine serum(Thermo Scientific, Northumberland, UK)], which was replaced daily.

After 2 to 3 months, under these conditions, hESC-RPE and hiPSC-RPE would form monolayer sheets of pigmented cells that can be dissected for gene expression analysis.

Small RNA isolation

The total RNA containing the miRNA species were extracted from pigment cluster and RPE cells using mirVana miRNA Isolation kit (Ambion). The RNA quality and yields were analyzed by gel electrophoresis and Nanodrop. RNA samples were then aliquoted and stored at -80°C for small RNA library construction.

Construction of small RNA library and high-throughput sequencing

Total RNA (1~10 μg) was ligated with a pair of Illumina adaptors to their 5’ and 3’ ends. After reverse transcription, the small RNA molecules were amplified using multiplex adapter primers for 12 cycles and the fragments around 95~100 bp (small RNA adapter) were then isolated from agarose gel. The purified DNA was used directly for cluster generation and sequencing analysis using the Illumina’s HiSeq2000 sequencing according to the manufacturer’s instructions. Then the image files generated by the sequencer were processed to produce digital-quality data. Reads were trimmed for the 3’ adapter, then mapped the hg19 genome using Bowtie [17]. We accepted only reads that had at most 1 mismatch to the hg19 genome and mapped to no more than 10 places in the genome; the exact parameters in Bowtie are -v1 -m10 --all -strata.

Statistical analysis and bioinformatics

Data analysis was carried out using Matlab. Briefly, differential expression analysis between four differentiation stages and different somatic tissues was determined using a combination of one-way ANOVA, false discovery rate, and fold change. Scripts are available upon request. The gene targets of differentially expressed miRNAs were predicted by two miRNA target prediction algorithms miRanda (http://microrna.sanger.ac.uk/sequences/), and Targetscan (http://www.targetscan.org/). We
used a common cutoff of mirSVR score $\leq 1.2$ as a threshold to reduce the fdr-true-positive rate of predicted targets. Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis of predicted target genes of miRNAs was performed by using DAVID 6.7, a web-based application (http://david.abcc.ncifcrf.gov/home.jsp) (default parameters). The obtained BP and KEGG categories were filtered for FDR<0.05 against the Homo Sapiens background. The relationship between candidate miRNA predicted to the affected mRNA and genes were constructed via NAVIGATOR software.

Quantification of mature miRNAs by real-time qRT-PCR

Real-time reverse transcription reactions containing 1–10 ng total RNA, 0.3 mM each dNTP, 5U MultiScribe Reverse Transcriptase, 1x RT buffer, 2.5 mM dNTP, 40 U RNSa RNAse Inhibitor and 1x TaqMan Reverse transcription reactions were incubated at 16°C for 30 min, 42°C for 30 min, 95°C for 5 min, then store at -80°C until use in TaqMan assays. 20 μl TaqMan real-time PCR reactions consist of 1x TaqMan Universal PCR Master Mix, 1x TaqMan mRNA assay, 1.5 μl of cDNA, and nuclease free water. Each TaqMan miRNA assay was done in duplicate for each sample tested. Relative quantities were calculated using the $2^{-\Delta\Delta Ct}$ method with U6 snRNA as endogenous control. Reactions were run with the Standard 7300 default cycling protocol without the 50°C incubation stage, with reactions incubated at 95°C 10 min, followed 40 cycles of 95°C 13 sec, 60°C 1 min. Fluorescence readings were collected during the 60°C 13 sec.

Acknowledgments

This work was supported by grants from the National Institutes of Health (R01 AG024889, R01 NS070018, R01 NS072116, P50 GM081897, T32 DC000024, KL2 TR000139, and T32 HL067748) and the American Lung Association, to AMM. The authors are also supported in part by NIH-NCATS CTSA UL1 TR000124.

Author Contributions

Conceived and designed the experiments: GKH JY GT ZX. Performed the experiments: GJY SPG JK JX ZL. Analyzed the data: GKH KZ. Contributed reagents/materials/analysis tools: GKH JY SPG JK JX ZL KJ JY YL SY GF LEI ZX. Wrote the paper: GKH GF ZX.
Chapter 7
Analysis of DNA methylation and its impact on RNA transcription during the differentiation of human pluripotent stem cells into retinal pigment epithelial cells

Zhenshan Liu\textsuperscript{1}, Rongfeng Jiang\textsuperscript{1}, Kevin Huang\textsuperscript{2}, Yun Feng\textsuperscript{1}, Ganlu Hu\textsuperscript{1}, He Xu\textsuperscript{2}, Songtao Yuan\textsuperscript{3}, Qinghuai Liu\textsuperscript{3}, Guotong Xu\textsuperscript{4*}, Zhigang Xue\textsuperscript{1*}, and Guoping Fan\textsuperscript{1, 2*}

\textsuperscript{1}Stem Cell Research Center, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai 200092, China
\textsuperscript{2}Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California 90095, USA
\textsuperscript{3}Department of Ophthalmology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu Province 210029, China
\textsuperscript{4}Tongji Eye Institute and Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, 200092, China
Abstract:
Using the paradigm of in vitro differentiation of hESCs/iPSCs into retinal pigment epithelial (RPE) cells, we have previously profiled mRNA and miRNA transcriptomes to define a set of RPE mRNA signature genes and a subset of miRNAs that are involved in RPE differentiation. In this study, to understand the role of DNA methylation in the differentiation, we profiled genome-scale DNA methylation patterns using the reduced representation bisulfite sequencing (RRBS) method. We found that waves of de novo methylation and demethylation took place in four designated stages of RPE differentiation. Integrated analysis of DNA methylation and RPE transcriptomes revealed a reverse-correlation between levels of DNA methylation and expression of a subset of miRNA and mRNA genes that are critical for RPE differentiation. Gene Ontology (GO) analysis suggested that those genes undergoing dynamic methylation changes were important for RPE differentiation and maturation. We further compared methylation patterns among human ESC- and iPSC-derived RPE as well as primary fetal RPE (fRPE) cells, and discovered that DNA methylation patterns could be used to classify specific stages of RPE cell differentiation. These results demonstrate that DNA methylation patterns may be useful biomarkers to characterize degrees of stem cell differentiation during the conversion of human pluripotent stem cells into functional RPE cells.
Background

DNA methylation is an important epigenetic modification involved in numerous cellular processes, including embryonic development [1-3], genomic imprinting [4, 5], X-chromosome inactivation [6, 7], and chromosome stability [8]. During development, DNA methylation plays an important role in epigenetic programming by silencing stem cell-specific genes and activating differentiation-associated genes [9, 10]. Recent studies using high-throughput sequencing technologies have mapped the genome-wide DNA methylation changes in single nucleotide resolution. These studies have uncovered that DNA methylation contributes to cellular lineage commitment in vitro [11-13] and in vivo [14-17].

The retinal pigment epithelium (RPE) is a monolayer of terminally differentiated pigmented cells between the neural retina and choroid. RPE is involved in the formation of the blood-retinal barrier, absorption of stray light, supplying of nutrients to the neural retina, regeneration of visual pigment, as well as the uptake and recycling of the outer segments of photoreceptors. We and other labs have achieved the ability to generate functional RPE through in vitro differentiation of both human embryonic stem cells (hESCs) and induced pluripotent stem cells (hiPSCs) [18-24]. Furthermore, RPE-derived from hESCs and hiPSCs can be injected into subretina space where normal RPE resides and restore visual function in the retinal dystrophy rat model [23, 25]. To understand gene regulation during in vitro differentiation of hESCs/iPSCs into RPE, we have previously identified RPE mRNA signature genes [20] and demonstrated that RPE-specific miRNAs were associated with the RPE differentiation and maturation of RPE in vitro [24].

In this study, we mapped genome-scale DNA methylation at single-base resolution during the course of RPE differentiation at four distinct stages. We further correlated promoter
DNA methylation with both mRNA and miRNA gene expression during in vitro RPE differentiation.

**Results**

**Profiling genome-scale DNA methylation patterns during the differentiation of human stem cells into RPE cells**

We have derived functional RPE cells from multiple lines of human pluripotent stem cells including total thirteen lines of hESCs and iPSCs through in vitro differentiation over the course of three to six months [20, 24] (data now shown). In our observations, we found that both H9 and UCLA4 hESCs as well as hiPSC2 and HDF2 iPSCs are representative of all hESCs and iPSCs in the RPE differentiation time course [20]. Moreover, both hESC-derived RPE and iPSC-derived RPE cells are well characterized in their cell biology profiles including RPE marker expression and RPE functional phagocytosis assays [20, 24] (data now shown). In this study, we focused on the characterization of DNA methylation profiles during directed RPE differentiation of two pairs of hESCs (H9 and UCLA4) and hiPSCs (hiPSC2 and HDF2) and cross-referenced methylation profiles with mRNA and miRNA expression profiles. We also chose two well characterized primary fetal RPE cells as control [20, 24]. Genomic DNA was isolated from four distinct stages of RPE differentiation including: 1) pluripotent stem cells (H9 and UCLA4 hESCs as well as hiPSC2 and HDF2 iPSCs); 2) 15 days partially differentiated ES cells; 3) early pigmented clusters after 30-day in vitro differentiation; and 4) functional RPE (3-6 months in culture). DNA methylation mapping was performed by the RRBS method, which is a robust, quantitative, and effective approach to map global DNA methylation. Our RRBS analyses
covered on average about 1 million individual CpGs throughout the human genome (see Supplementary Table 1), including those found within over ten thousand unique gene promoters.

To assess whether DNA methylation patterns distinguish cell types, we performed hierarchical clustering and principal component analysis (PCA) based on genome-wide CG methylation levels (Figure 7-1). Both clustering methods revealed that terminally differentiated cells clustered distinctly away from immature cell types, such as ESCs and iPSCs and partially differentiated cells. Furthermore, we observed ESC-derived RPE (ESC-RPE) were more similar to each other than iPSC-derived RPE (iPSC-RPE). However, both ESC-RPE and iPSC-RPE were distinctly different to fetal RPE. Overall, identical cell types clustered tightly together, suggesting that each cell-type exhibits a well-defined DNA methylation pattern. These observations were reminiscent of what we have observed with mRNA and miRNA expression patterns [20, 24], suggesting that DNA methylation patterns are as useful as gene expression profiles to distinguish cell types during RPE differentiation.

**DNA methylation dynamics during RPE differentiation**

We next examined the DNA methylation dynamics at four distinct stages of RPE differentiation (Figure 7-2A). We found that the level of CpG methylation was initially increased upon partial stem cell differentiation, then decreased during lineage-specification, and then increased again upon RPE maturation (Figure 7-2B). In addition to the difference in overall methylation, we also analyzed the distribution of the CpG methylation levels in the four RPE differentiation stages (Figure S1). These findings indicate that both de novo methylation and demethylation dynamically take place during RPE differentiation.
The first wave of methylation changes from hESCs (Stage I) to partially differentiated ES cells (PD, Stage II) were dominated by hypermethylation. However, transition from Stage II partially differentiated ES cells to Stage III pigmented clusters showed more hypomethylation than hypermethylation (Figure 7-2B). To understand which genomic elements undergo DNA methylation changes during RPE differentiation, we next analyzed the distribution of hyper- and hypo-methylated CpG sites between two adjacent stages. The results showed that while repeat elements [long interspersed nuclear element (LINE), long terminal repeat (LTR), and satellite repeats] and gene bodies did not exhibit much difference, a subset of CpG islands (CGI) and short interspersed nuclear elements (SINEs) showed differential methylation during RPE differentiation (Figure 7-2C and D). Further analysis of these genomic elements indicated that the distribution of methylation levels were quite consistent with the DNA methylation dynamics at four distinct stages of RPE differentiation (Figure 7-2C and D, Figure S3).

Changes in promoter methylation during RPE differentiation

Recent studies highlight that DNA methylation patterns change significantly during the differentiation processes such as from pluripotent stem cells or multipotent progenitors to lineage-committed cells [11, 26, 27]. In RPE differentiation (Figure 7-3A), a subset of genes associated with cell membrane function are dynamically methylated from hESCs to PD Stage II cells, and to Stage III PC cells. In the transition stage between Stage II PD cells and Stage III PC cells, demethylated genes were enriched in metabolic and biosynthetic process (Figure 7-3B). In contrast, a subset of genes associated with non-membrane-bounded organelle and transcription factor activity exhibits increased methylation (data not shown). Finally, during RPE maturation (Stage III to Stage IV), we found that these genes are associated with ribonucleotide and
nucleotide binding (Figure 7-3C). GO annotations for all re-methylated genes showed significant enrichments for genes in the developmental processes and transcriptional regulation (Figure 7-3D), consistent with the silencing of these developmental genes in the terminally differentiated RPE cells, including LAMC2, PIM2, and CD74 (Figure S4) [28]. Overall, these data indicate that locus-specific DNA methylation patterns may be useful for identifying developmentally regulated RPE-differentiation genes.

**DNA methylation and miRNA expression changes during RPE differentiation**

MicroRNAs are small non-coding RNAs which are expressed in a tissue-specific manner and play important roles in cell proliferation and differentiation. DNA methylation-mediated downregulation of miRNA gene expression has been observed in various types of cancers [29]. Since we have recently profiled the miRNA expression pattern during RPE differentiation [24], we further analyzed the data to determine any correlation between DNA methylation and miRNA expression. In total, we analyzed 419 miRNAs for their expression and DNA methylation levels, and selected 216 high and 92 low methylated miRNAs based on H9 RPE methylation levels (methylation levels >0.8 were defined as high, and methylation levels <0.2 were defined as low, see the supplement Figure S5). In mature RPE, we found a negative correlation between DNA methylation and miRNA expression for hsa-mir-193b and hsa-mir-210 clusters in the four different cell stages (Figure 7-4), suggesting these two miRNAs may be directly regulated by DNA methylation. However, we also found 18 cases where DNA methylation and miRNA expression were not anti-correlated (e.g., has-mir-181c in Figure 4A), suggesting DNA methylation-independent regulation for these microRNAs.
Most RPE mRNA signature genes are demethylated at proximal promoter regions in hESC-derived RPE

DNA methylation is linked to gene silencing and considered to be an important mechanism in the regulation of mRNA transcription. Previously we have identified a set of 87 RPE signature genes in fetal and stem cell-derived RPE [20]. To examine the relationship between DNA demethylation and gene expression of these RPE mRNA signature genes, we determined the DNA methylation levels of promoters in pluripotent stem cells and RPE. Surprisingly, among the 46 RPE mRNA signature genes detected via RRBS method, we found most of these genes maintained proximal promoter hypomethylation in both stem cells and RPE stages. Our observations suggest that either: 1) gene coverage via RRBS is incomplete, thus inadequate to predict the status of other signature genes; or 2) demethylated promoters are permissive to be expressed, but only activated via other regulatory mechanisms in mature RPE cells. Nevertheless, two RPE signature genes, G protein-coupled receptor 143 (GPR143) and chloride channel (CLCN4) exhibited promoter demethylation coupled with gene activation from undifferentiated and partially differentiated stem cells to pigmented cluster and RPE cells, implicating that DNA demethylation contributes the activation of these two RPE mRNA signature genes during differentiation (Figure 7-5A). Next, we analyzed the DNA methylation of 26 ES-cell specific genes during RPE differentiation [30-32]. We found that the methylation levels in a portion of these genes tend to be hypermethylated and silenced in differentiated RPE cells, such as DPPA2, TDGF1 and SALL4 (Figure 7-5B).

Fetal RPE show hypomethylation in cell adhesion and extracellular matrix genes
Recent studies found that hESC-RPEs and iPSC-RPEs expressed essential RPE markers and could rescue visual function in animal models, and have the potential to treat a wide range of retinal diseases [18, 19, 23, 33]. Our previous study found that despite morphology and functional similarities, gene expression of stem-cell-derived RPE cells were moderately different from normal fRPE cells, and that fRPE-specific genes were important for eye development [20]. To further determine whether hESC-RPEs and hiPSC-RPEs mimic normal RPE cells, we asked if these stem-cell-derived RPE cells shared similar DNA methylation profiles. In our analyses, we noticed that there was an obvious variability in DNA methylation patterns between hESC-RPEs, iPSC-RPEs, and fetal RPE (Figure 7-6A). We found that fetal RPE exhibited a set of demethylated genes that were highly methylated in both hESC-RPE and hiPSC-RPE. GO analysis revealed that these genes were associated with cell adhesion and ion binding (Figure 7-6B). These differentially methylated genes were modestly inversely-correlated with gene expression (Figure 7-7). We also found that hESC-RPEs and iPSC-RPEs had slightly different DNA methylation profiles from each other (e.g., 20 differentially methylated genes as in Figure 7-6A). GO analysis revealed that these genes were associated with peptidase activity (data not shown). These results demonstrate that significant differences in DNA methylation profiles exist among hESC-RPE, iPSC-RPE, and fRPE. Many of these CpG methylation patterns sites were already established in undifferentiated hESCs and iPSCs (Figure 7-6A). Furthermore, some differential methylation between hESC-derived RPE and iPSC-derived RPE is reflected in previous differential methylation found between undifferentiated hESCs and iPSCs (as illustrated in the bottom quarter of Figure 7-6A).

Discussion
Gene expression profiles have been shown to accurately reflect cell type and differentiation stage [16, 34, 35]. In this study, we found that four stages of RPE differentiation can be distinguished through a unique subset of DNA methylation patterns. Although previous studies have shown that iPSCs and ESCs are similar in terms of transcription program, chromatin modification profiles [37-42], and global chromatin configuration [37, 43, 44], differences in epigenomics [45-47] and differentiation potential between iPSCs and ESCs [48-50] have been previously reported. In this study, by establishing high-resolution genomic maps of DNA methylation for stem-cell-derived RPE and human fRPE cells, we revealed the distinct differences in DNA methylation among hESC-RPE, iPSC-RPE, and fRPE. Our findings also support the idea that hESCs and iPSCs may have inherent differences in DNA methylation patterns [41, 42, 45].

Previous studies have revealed that promoter DNA methylation has an important effect on regulating cell type specific genes, which ultimately contributes to both cell physiology and morphology [51, 52]. Ji et al. found that differential DNA methylation is more strongly correlated with gene expression at CpG island shores than CpG islands [15]. Even intragenic methylation could have an important role in regulating cell context-specific alternative promoters in gene bodies [53]. However, other studies argued that DNA methylation might only exert a minor influence on the regulation of tissue-specific gene expression [54]. In this study, for some genes (e.g., GPR143, CLCN4, CHCHD2, SLC16A8, hsa-mir-193b and hsa-mir-210), we found the negative correlation between promoter DNA methylation and gene expression, consistent with the hypothesis that methylation represses gene expression. Nevertheless, for many RPE signature genes, DNA methylation and gene expression did not show strong anti-correlation during the RPE differentiation process. One possible explanation for this observation is that our RRBS assay is limited in gene coverage to reveal differential methylation patterns. Alternatively,
our results suggest that other regulatory mechanism(s) besides DNA methylation is more important for regulating RPE-specific genes. Future study with comprehensive profiling of genome-wide DNA methylation and histone modifications may shed additional insights on the relationship between methylation regulation and RPE gene expression during cell differentiation.

Dysfunction, degeneration, and loss of RPE cells are prominent features of Best disease, subtypes of retinitis pigmentosa (RP), and age-related macular degeneration (AMD). Because current treatments for these diseases are severely limited, stem cell-based replacement therapy involving RPE transplantation holds tremendous promise. The human pluripotent stem cells (hESC and hiPSC) may serve as an unlimited donor source of RPE cells for transplantation. Previous studies on differentiating RPE cells from stem cells demonstrated that stem-cell-derived RPE cells had molecular characteristics similar to primary RPE cells [18, 19, 33]. Moreover, the transplantation of stem-cell-derived RPE could partially restore visual function in the retinal dystrophy rat model [23, 33, 55]. To assess the potential of stem cell-derived RPE for cell replacement therapy, Sugino et al compared the attachment and survival of hES-RPE of different degrees of pigmentation on Bruch's Membrane (BM) with cultured human fRPE, and found that hES-RPE showed impaired initial attachment, and cell behavior and protein secretion were markedly dissimilar [56]. Of note, in the current study, we found that a bigger difference in DNA methylation between fRPE and stem cell-derived RPE exists for those genes functioning in cell adhesion and ion transport. It will be of interests to understand whether differential methylation leads to any functional alteration in RPE cell adhesion in the above cell transplantation studies.

Conclusions
To our knowledge, this is the first study that has used a genome-wide DNA methylation approach to study the role of DNA methylation during in vitro differentiation of RPE from human pluripotent stem cells. We found that DNA methylation is dynamic during RPE differentiation, and found an obvious variability among human ESC- and iPSC-derived RPE as well as primary fRPE cells. These observations demonstrate that DNA methylation could accurately reflect cellular identity and distinguish different stages during RPE differentiation. Integrated analysis of DNA methylation and RPE transcriptome revealed a reverse-correlation between levels of DNA methylation and expression of a subset of miRNA and mRNA-coding genes that are critical for RPE differentiation, suggesting that RPE-specific differentially methylated genes play an important role during RPE differentiation.

Materials & Methods

Differentiation of hESC/hiPSC-RPE

The human embryonic stem cell lines H9 and UCLA4, induced pluripotent stem cell lines hiPSC2 and HDF iPSC were obtained from the UCLA Stem Cell Core. Fetal RPE cell lines fRPE1 and fRPE2, and mouse embryonic feeder cells are generated in Dr. Guoping Fan’s lab at UCLA. Pluripotent stem cells (hESC and iPSCs) were plated onto gamma-rays irradiation mouse embryonic feeder cells with DMEM/F12 culture medium containing 20% Knock-Out Serum Replacement, 0.1 mM nonessential amino acids, 0.1 mM β-mercaptoethanol and 100 ng/ml zebrafish basic fibroblast growth factor (zfbFGF) on a 6-well plate. Briefly, cells were cultured at 37°C in 5% CO₂ for 6–10 days after which zfbFGF was omitted to facilitate spontaneous cell differentiation.
Pigmented colonies were observed within 4 weeks and allowed to expand for a few weeks, with media changes every 2–3 days. Pigmented cells were enriched by manual dissection using insulin needle followed by seeded on growth factor reduced Matrigel (BD Biosciences, diluted 1:30) coated plate and transwell membranes.

RPE medium were changed to support pigment cluster expansion [containing α-MEM, 1×N2 supplement (Gibco), 1×Non-essential amino acid solution, 250 mg/ml taurine, 13 ng/ml Triiodothyronin (Sigma-Aldrich, Gillingham, UK), 20 ng/ml Hydrocortisone (Sigma), 2 mM L-glutamine (Invitrogen, Paisley, UK), 1×Penicillin-streptomycin and 10% Hyclone heat-inactivated fetal bovine serum (Thermo Scientific, Northumberland, UK)], which was replaced daily.

**DNA Isolation**
Genomic DNA was isolated from all samples by traditional phenol/chloroform method. DNA quality was controlled by agarose gel electrophoresis and quantified by a NanoDrop ND-1000 Spectrometer (PeqLab Biotechnologies, Erlangen, Germany).

**Reduced representation bisulfite sequencing**
Reduced representation bisulfite sequencing (RRBS) was performed as described [57]. RRBS reads cover less than 10% of the 28 million CpGs in the human genome [58]. Briefly, 1 µg genomic DNA was digested with the methylation insensitive restriction enzyme MspI (NEB). Ends of each restriction fragment were filled in and a 3′ adenosine was added with Klenow Fragment (3′→5′ exo-minus, NEB). Methylated paired-end Illumina adapters were ligated to the ends of the DNA fragments using T4 DNA Ligase (NEB). Fragments between 100 bp and 400 bp were purified by agarose gel extraction. The purified fragments were treated with sodium
bisulfite and then amplified by PCR. The final PCR products were sequenced on Illumina HiSeq2000 machines.

**Statistical analysis and bioinformatics**

The RefSeq Genes NCBI36/hg18 was downloaded from UCSC Genome Browser (http://genome.ucsc.edu/). A masked hg18 genome was generated by in silico digestion of the genome, followed by masking all regions that were not within theoretical range of MspI digestion and expected size selection. Reads were mapped to this RRBS genome using BS-Seeker [59], and methylation calling was performed as previously described [11]. Individual cytosine methylation levels was defined as the number of unconverted cytosine to number of unconverted and converted cytosine [\#C/(\#T+\#C)]. Only cytosines that were covered at least 5 reads were considered for further downstream analysis. Clustering analysis and heatmaps were performed using built-in functions in Matlab. Only CGs that were found in all cell types assayed were considered. For DNA methylation heatmaps, colors represent average methylation levels ranging between 0-100%.

For successive stage changes in CG methylation, we examined the number of shared CGs that changed by >50%, which are significantly differentially methylated as tested by binomial cumulative distribution function. For promoters, we considered only CGs that were -1000 bp to +500 bp of the transcription start site (TSS), and required differential methylation difference of >40% and p-value < 0.05 as calculated by Student T-test. Genomic distribution of CGs were performed using annotateBed found in the BEDTools package[60]. GO analysis was performed using DAVID [61].

**Accession IDs**
RRBS data reported in this paper have been deposited in the NCBI Gene Expression Omnibus (GEO) database with accession number GSE43473.

**Abbreviations**

hESCs: human embryonic stem cells; iPSCs: induced pluripotent stem cells; RPE: retinal pigment epithelial; RRBS: reduced representation bisulfite sequencing; GO: Gene Ontology; fRPE: fetal RPE; PCA: principal component analysis; PD: partially differentiated ES cells; PC: pigmented clusters; ESC-RPE: ESC-derived RPE; iPSC-RPE: iPSC-derived RPE; RP: retinitis pigmentosa; AMD: age-related macular degeneration; BM: Bruch's Membrane; zfbFGF: zebrafish basic fibroblast growth factor; bp: base pair; TSS: transcription start site; LINE: long interspersed nuclear element; LTR: long terminal repeat; SINEs: short interspersed nuclear elements.

**Acknowledgements**

This work was supported by NIH grant GM081621, International Science & Technology Cooperation Program of China (2011DFB30010), National Natural Science Foundation of China (81271258), Natural Science Foundation of Jiangsu Province of China (DK2011321), and Ministry of Science and Technology 973 Basic Research program in China (2011CB965102, 2011CB966204 and 2012CB966303).

**Authors contributions**

Zhenshan Liu, Yun Feng, He Xu, and Ganlu Hu carried out the experiments including hESC/hiPSC culture, RPE differentiation and RRBS libraries construction. Rongfeng Jiang and
Kevin Huang contributed to bioinformatics and statistics. Zhenshan Liu and Guoping Fan
designed the study and wrote the manuscript. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.
References


9. Kashyap V, Rezende NC, Scotland KB, Shaffer SM, Persson JL, Gudas LJ, Mongan NP: Regulation of stem cell pluripotency and differentiation involves a mutual regulatory circuit of the NANOG, OCT4, and SOX2 pluripotency transcription factors with polycomb


Derivation of functional retinal pigmented epithelium from induced pluripotent stem cells.
DS et al: Molecular signature of primary retinal pigment epithelium and stem-cell-derived
pigment epithelium (RPE) cells exhibit ion transport, membrane potential, polarized
vascular endothelial growth factor secretion, and gene expression pattern similar to native
22. Lamba DA, Karl MO, Ware CB, Reh TA: Efficient generation of retinal progenitor cells
from human embryonic stem cells. Proceedings of the National Academy of Sciences of the
N, Sauve Y, Lanza R: Human embryonic stem cell-derived cells rescue visual function in


Figure legends

Figure 7-1. Different cell types have distinctive global DNA methylation signatures.
DNA methylation profiles of different cell types were clustered using either unbiased
hierarchical clustering or principal component analysis (PCA) based on the DNA methylation
levels of all shared CGs in all cell lines analyzed. A) Hierarchal clustering using Pearson
correlation distance between methylation levels of 535,376 shared CGs. Defined cell types show
a high overall similarity in the methylation patterns and thus cluster together. B) 2D-biplot of the
first two principal components. For this analysis, we report the average principal component
scores after randomly sampling of 100,000 shared CGs over 200 iterations.

Figure 7-2. DNA methylation dynamics during RPE differentiation.
(A) Bar graph displaying the mean CG methylation level of all assayed CGs that are shared
between the H9 line of RPE (N=733,672). (B) Bar graph showing the proportion of hyper- and
hypo-methylated CpG sites between two adjacent stages during RPE differentiation (differential
methylation > 50%). (C, D) The metaplots of dynamic CG methylation changes for (C) CGI and
(D) SINE during RPE differentiation.

Figure 7-3. DNA methylation at RPE-specific genes during RPE differentiation from
human ES cells.
(A) Heatmap analysis of differentially methylated genes during RPE differentiation. ESC is
undifferentiated H9 hESCs, PD: partial differentiated H9 cells, PC: pigmented cluster. RPE: H9-
derived RPE cells. (B, C, D) GO analysis of RPE-specific signature genes during RPE
differentiation. Bar graphs showing significance of enrichment terms for set of demethylated
genes from PD into PC cells (B) and during the course of RPE maturation from PC to RPE (C),
and remethylated genes in mature RPE (D) as listed in Supplementary Table 2-4. P-values < 0.05

Figure 7-4. DNA methylation and miRNA expression changes during RPE differentiation.
(A) Heatmap analysis of differentially expressed miRNAs during the course of differentiating H9
ESCs. miRNA expression levels were standardized by the mean level in all stages, whereas DNA
methylation levels not standardized. DNA methylation values were the average levels between -
1000 bp upstream and +500 bp downstream of the TSS. (B) Genome browser views of DNA
methylation patterns found in hsa-mir-193b and hsa-mir-210 loci during the course of RPE differentiation.

**Figure 7-5. DNA methylation profiles of RPE and ESC signature genes during the course of RPE differentiation**
Heatmap showing the average promoter methylation levels of (A) RPE and (B) signature genes during the course of RPE differentiation.

**Figure 7-6. Differential DNA methylation in hES/hiPSC-derived RPEs and fRPEs**
(A) Heatmap analysis of differential methylated genes in hES/hiPSC-derived RPEs and fRPEs. (B) GO analysis of fRPE-specific demethylated genes.

**Figure 7-7. The correlation of gene expression and DNA methylation changes in hES/hiPSC-derived RPEs and fRPEs**
(A) Scatter plot of gene expression fold change vs DNA methylation change in hES/hiPSC-derived RPEs and fRPEs. (B) Gene expression heatmap analysis of selected genes that are methylated in H9-RPE, but not in fetal RPE.
Figure 7-1.
Figure 7-2.
Figure 7-3.

A

B

C

D

GO Term

Intracellular signaling cascade
Positive regulation of nitrogen compound metabolic process
Lymphocyte differentiation
Positive regulation of cellular biosynthetic process
Positive regulation of biosynthetic process
Cytokine binding
Zinc ion binding
Phosphorus metabolic process
Phosphate metabolic process
Leukocyte differentiation

GO Term

Ribonucleotide binding
Purine ribonucleotide binding
Purine nucleotide binding

GO Term

Extracellular region part
Proteaceous extracellular matrix
Urogenital system development
Positive regulation of nitrogen compound metabolic process
Extracellular matrix
Tissue morphogenesis
Regulation of transcription, DNA-dependent
Positive regulation of cellular biosynthetic process
Extracellular region
Positive regulation of biosynthetic process
Skeletal system development
Positive regulation of nucleoside, nucleotide, nucleotide and nucleic acid metabolic process

-Log(P-value)

-Log(P-value)
Figure 7.4.

A

miRNA expression

DNA methylation

B

H9ES
H9PD
H9PC
H9RPE

chr16: 14,305,200 14,305,300 14,305,400 14,305,500

chr11: 558,000 558,500 559,000 559,500 560,000
Figure 7-5.
Figure 7-7.

A

B
Chapter 8
Genetic Programs in Human and Mouse Early Embryos Revealed by Single-Cell RNA-Sequencing

Zhigang Xue1*, Kevin Huang2*, Chaochao Cai2, Lingbo Cai3, Chun-yan Jiang3, Feng Yun1, Zhenshan Liu1, Qiao Zeng1, Liming Cheng1, Yi E. Sun1, Jia-yin Liu3*, Steve Horvath2, and Guoping Fan2*

1Translational Center for Stem Cell Research, Tongji Hospital, Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, 200065, China
2Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA
3State Key Laboratory of Reproductive Medicine, Center of Clinical Reproductive Medicine, First Affiliated Hospital, Nanjing Medical University, Nanjing 210029, China

*These authors contributed equally to this project.
Mammalian preimplantation development is a complex process involving dramatic changes in the transcriptional architecture\(^1^\text{-}^4\). Through single-cell RNA-sequencing (RNA-seq), we report here a comprehensive analysis of transcriptome dynamics from oocyte to morula in both human and mouse embryos. Based on single nucleotide variants (SNVs) in blastomere mRNAs and paternal-specific SNPs, we identify novel stage-specific monoallelic expression patterns for a significant portion of polymorphic gene transcripts (25-53\%). By weighted gene co-expression network analysis (WGCNA)\(^5^,^6\), we find that each developmental stage can be concisely delineated by a small number of functional modules of co-expressed genes. This result indicates a sequential order of transcriptional changes in pathways of cell cycle, gene regulation, translation, and metabolism in a step-wise fashion from cleavage to morula. Cross-species comparisons reveal that the majority of human stage-specific modules (7 out of 9) are remarkably preserved, only to diverge in developmental specificity and timing in mice. We further identify conserved key members (or hub genes) of the human and mouse networks. These genes represent novel candidates that are likely key players in driving mammalian preimplantation development. Collectively, we demonstrate that mammalian preimplantation development is orchestrated by evolutionarily conserved genetic programs that diverge in developmental timing. Our results provide a valuable resource to dissect gene regulatory mechanism underlying progressive development of early mammalian embryos.

Mammalian preimplantation development involves sequential decay of maternally stored RNAs in parallel with massive induction of new transcripts from the embryonic genome in a process called zygotic (or embryonic) genome activation (ZGA or EGA)\(^7^\text{-}^9\). However, previous gene expression profiling of human preimplantation have been limited by blastomere sample size
and quantitation platforms\textsuperscript{10-13}. Recent advances in single-cell RNA-seq technology\textsuperscript{14} have provided the unprecedented opportunity to study gene regulation in early human embryos under high resolution. We mapped ~700 million sequencing reads derived from 33 single cells of multiple stages, ranging from mature oocytes to 8-cell blastomeres, and 3 morula embryos (Supplementary Text and Supplementary Table 1). The typical number of detectable genes (RPKM $> 0.1$) ranged between $\sim 8,500$ to $\sim 12,000$ genes in individual cells, demonstrating our technique has greater sensitivity and coverage compared with previous microarray methods\textsuperscript{10-13}.

Unsupervised hierarchal clustering analysis and principal component analysis revealed that cells of different preimplantation stages form distinct clusters (Fig. 8-1a-b). Furthermore, single-cell resolution revealed that blastomeres from the same 8-cell embryo are more similar to each other than blastomeres from a separate 8-cell embryo (Fig. 8-1a-b).

Intriguingly, cluster analyses showed that both pronuclear and zygote (1-cell stage) embryos clustered closely together but away from oocytes and cleavage embryos, indicating that the 1-cell stage exhibits a distinct transcriptome pattern. Comparisons between mature oocytes and 1-cell embryos identified 149 differentially expressed genes (FDR $< 5\%$, $>$ 2 fold change), 79 and 70 of which were up- and down-regulated in 1-cell embryos, respectively (Fig. 8-1c). Interestingly, approximately half of the upregulated genes remain highly expressed in 2- and 4-cell stages but are downregulated at the 8-cell stage, whereas another half is further up-regulated at the 8-cell stage (Fig. 8-1d). By performing similar single-cell RNA-seq in mouse preimplantation embryos (Supplementary Table 1), we observed 520 transcripts that were up-regulated in pronuclei compared to mature oocytes (Supplementary Fig. 1), confirming that both human and mouse species exhibit a conserved minor wave of ZGA prior to the major wave
of ZGA (or EGA). These results indicated that our approach has overall greater quantitation for identifying gene expression changes between different stages of preimplantation development.

Single-cell RNA-seq enabled base-resolution scrutiny without confounding effects from cell population heterogeneity. In this unique study, all embryos were derived from intra- cytoplasmic sperm injection of different egg donors with the same sperm donor. Based on the paternal genotype as assayed by exome sequencing of the sperm donor’s blood sample, we were able to follow the paternal genome in embryos via single nucleotide polymorphism (SNP) analysis\textsuperscript{15-17}. We identified paternally or maternally expressed genes either by inferring phased paternal haplotype information\textsuperscript{18,19} or by leveraging sites which are paternally homozygous in exome sequencing, but heterozygous in embryo transcripts (Fig. 8-2a and Supplementary Table 2). In total, we determined the parent-of-origin expression for approximately 850-1,400 gene transcripts (or 15-20\% of all detected gene transcripts per stage). Surprisingly, we found that 53\% of 8-cell embryos transcripts and 23\% of morula transcripts exhibit monoallelic maternal expression patterns (Fig. 8-2b). For instance, within a single haplotype region, we observed simultaneous maternal activation and degradation in the ASB6 locus and C9orf78 locus, respectively (Fig. 8-2c). In a different scenario, the paternal genotype is homozygous in 3 consecutive SNPs (rs3829009, rs6990278, and rs8537) at the cell-cycle regulator CDCA2 gene locus. In this example, the phased paternal SNP pattern is not detected in 2- and 4-cell embryos but seen in 8-cell embryos, indicating that the detected transcripts in 2- and 4-cell embryos must be of maternal origin and also that CDCA2 undergoes paternal activation at the 8-cell stage (Fig. 8-2d). However, paternal activation at this locus appears transient as the paternal allele could not be detected in the morula stage. Significantly, we observed a moderate correlation between expression of the paternal allele and overall transcript expression (r=0.53, p=0.06), suggesting
the paternal allele is dynamically regulated. As further validation of allele-specific expression in early embryos, we analyzed the status of imprinting genes that exhibit known parent-of-origin expression patterns (see Supplementary Text).

The above analyses focused on individual gene transcript changes during each transitional stage, but did not reveal the crucial shift of gene networks in preimplantation development. To understand the co-expression relationships between genes at a systems level, we performed weighted gene co-expression network analysis (WGCNA)\(^5,6\). This unsupervised and unbiased analysis identified distinct co-expression modules corresponding to clusters of correlated transcripts (Fig. 8-3a). Strikingly, 9 out of 25 co-expression modules showed stage specific expression values, i.e., these modules are comprised of genes that tend to be over-expressed in a single developmental stage (\(r >0.7, \ p<10^{-3}\), Supplementary Fig. 2).

These stage-specific modules likely represent core gene networks operating in each transitional stage. In total, 8,301 genes were part of human oocyte to morula stage-specific modules revealing a step-wise requirement of new transcripts that are involved in gene transcription (4-cell), post-transcriptional RNA processing (8-cell), and then protein translation and cell energetics (morula) (Fig. 8-3b). As expected, the oocyte module undergoes gradual degradation over the course of development, whereas modules in the 4- to 8-cell transition show sharp degradation (3-fold) and activation (4-fold) (Fig. 8-3c). Collectively, our systems analysis revealed that the transcriptional organization for human preimplantation development can be summarized using a small number of stage-specific modules with well-defined function.

Significantly, these modules are highly robust and reproducible as can be seen from our module preservation analysis\(^20\) (Fig. 8-4a). This module preservation analysis uses a permutation test to define a test statistic, \(Z_{\text{summary}}\), which summarizes the evidence that the network topology
of the module is preserved in an independent dataset (see Methods). Seven out of 9 modules were detected with highly significant preservation scores ($Z_{summary} > 10$) in two publically available human preimplantation datasets (though with significantly less RNA transcripts). Prior to the direct study of transcription changes during human preimplantation process, mouse models have provided crucial insights into the preimplantation gene regulatory network\textsuperscript{2-4}. Apparently, only one study (involving far fewer transcripts) performed a cross-species comparison in the context of a pluripotency network\textsuperscript{11}. Therefore, we performed single-cell RNA-seq in mouse blastomeres from oocyte to morula embryos for a direct cross-species comparative analysis (Supplementary Table 1). Cross-species module preservation analysis\textsuperscript{20} showed that 7 out of 9 human stage-specific modules were at least moderately preserved ($Z_{summary} > 5$) (Fig. 8-4a and Supplementary Table 1). We further validated the cross-species preservation using other mouse preimplantation datasets\textsuperscript{11,21}, which confirmed most human-stage specific modules are preserved in mouse (Fig. 8-4a). As expected, intra-species module preservation was more significant ($Z_{summary} > 10$) than inter-species preservation ($Z_{summary} > 5$, Fig. 8-4a). Thus, these preserved gene networks represent a strongly conserved transcriptional architecture of key developmental programs.

By independently applying WGCNA to our mouse dataset, we found that mouse development also involves stage-specific co-expression modules (Supplementary Fig. 3). Gene ontology analyses showed that mouse stage-specific modules share many functional similarities to their human counterparts, including the conserved sequential activation of functional enrichment changes (Supplementary Fig. 4 and 5). As expected, overlaps between human and mouse stage-specific modules were highly significant ($p<10^{-4}$, Fig. 8-4b), and the functional enrichment of overlapped stage-specific genes mostly reflected the overall network function (Fig.
However, when we examined overlap of pre-major ZGA in human and mouse, we found that most genes were enriched for protein transport and GTPase signaling, whereas their respective modules as a whole were more enriched for cell-cycle genes (Fig. 8-4c and Supplementary Figure 4).

Interestingly, stage-specific modules in human and mouse overlapped across multiple stages. For example, the mouse oocyte and 1-cell module genes significantly overlapped with genes that were specific to human oocyte, 1-cell, and 4-cell stages ($p<10^{-6}$, Fig. 8-4b). This result suggested that mouse pre-major ZGA genes are spread over the longer gestational pre-major EGA window in humans. Likewise, post-major ZGA modules in mouse were found to have significant overlap and spread throughout all post-major EGA human stages (Fig. 8-4c).

Although the major ZGA diverges in timing between human and mouse, these two species re-converge in both timing and function at the 8-cell to morula transition, during which many mitochondrion related transcripts are up-regulated (Fig. 4c and Supplementary Fig. 4).

Collectively, our results show that human and mouse share many core transcriptional programs in early development, but diverge in their stage-specificity and timing, likely reflecting species-specific differences in human and mouse gestational periods.

Using the WGCNA measure of intramodular gene connectivity (kME), we identified 491 intramodular hub genes across all stage-specific modules (kME>0.9, $p<10^{-22}$). Intramodular hub genes are centrally located in their respective modules and may thus be critical components within the network. Remarkably, 337 (69%, $p<10^{-16}$) intramodular hub genes in stage-specific modules can be validated (i.e. independently identified as a hub gene in a separate dataset), demonstrating that preimplantation hub genes are highly reproducible (Supplementary Table 3). For example, KPNA7 (kME=0.89, $p<10^{-22}$) is consistently identified as a hub gene in multiple
human and mouse pre-major EGA networks (kME>0.91). Significantly, Kpna7-deficient embryos have previously been shown to arrest at 2-cell stage and fail to activate zygotic genes\textsuperscript{22}. Furthermore, SIN3A, a transcriptional co-repressor, is a 4-cell intramodular hub gene that was validated in every independent dataset (Fig. 8-4d and Supplementary Table 3). Indeed, analysis of upstream regulators of 8-cell module genes revealed overrepresentation in MYC, MAX, and MXI1 targets, which are absent in the 4-cell stage modules. This finding is consistent with the previously characterized activation of Myc enriched genes upon Sin3a depletion\textsuperscript{23}. Together, these results demonstrate that some intramodular hubs are likely key players of preimplantation development.

In summary, we demonstrated that single cell RNA-sequencing has significantly improved transcriptome quantification of scarce human preimplantation embryo samples at both individual transcript and system levels. Our findings significantly extend knowledge of the transcriptional architecture, sequential order of gene activation, and genetic programming for early human embryogenesis. Our cross-species systems analysis demonstrated that the preimplantation transcriptional organization is highly conserved, highlighting an evolutionarily conserved molecular process including key genes that drive mammalian preimplantation development. It remains to be resolved whether the differences in a portion of co-expressed modules can account for species-specific function, such as previously seen in systems comparisons of human and mouse brain\textsuperscript{24}. Furthermore, we expect that single cell RNA-seq can also quantitatively delineate the structures, isoforms, and allele-specific expression patterns of both coding genes and non-coding regulatory RNAs\textsuperscript{8,13,21}. Thus, this technology is well suited for revealing the exquisite nature of genetic programming in early mammalian embryogenesis. Compared to exome/genomic sequencing, RNA-seq analysis has the advantage to quantitatively
reveal expression defects due to genetic and/or epigenetic alternations in gene regulatory domains that are beyond the detection of DNA sequencing, such as in the case of imprinting disorders. Thus, single-cell RNA-seq of a single blastomere could be a promising approach for preimplantation genetic diagnosis in the near future.

METHODS SUMMARY

Oocytes and early embryos were obtained from the Center for Clinical Reproductive Medicine at the Jiangsu People’s Hospital, 1st Affiliated Hospital of Nanjing Medical University with patient’s written informed consent and institutional approval. Single cell isolation and RNA-Seq experiments were performed as described\textsuperscript{14} with use of the Illumina HiSeq2000 instrument according to manufacturer specifications. Detailed information on ethical conduct, sample collection, and data analyses are available in Methods.

Acknowledgements We thank many of our colleagues for invaluable discussions and comments on our study. Paivi Pajukanta at UCLA provided very helpful suggestions on SNP analyses and Dan Geschwind for critically reading the manuscript. This work was supported by 973 Grant Program #2012CB966300, 2011CB966204, and 2011CB965102 from Ministry of Science and Technology in China; the International Science & Technology Cooperation Program of China (No.2011DFB30010); National Natural Science Foundation of China (81271258).

Author contributions ZX, KH, JY, and GF designed the study. ZX, LC, CY, FY, ZL, LC, YS performed experiments or contributed critical reagents and protocols. KH, CC, and SH analyzed
the data and performed statistical analyses. KH and GF wrote the manuscript in discussion with all the authors. All the authors read and approved the manuscript.

**Author information** All sequencing data generated for this work have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession number **GSE44183**. The authors declare no competing financial interests. Correspondence and request for materials should be addressed to GF (gfan@mednet.ucla.edu), JL (jyliu_nj@126.com), and ZX (xuezhigang75@gmail.com).
**Figure Legends**

**Fig. 1. High resolution single-cell transcriptome analysis of human preimplantation embryos.** a) Unsupervised hierarchical clustering and b) principal component analysis of single blastomere expression patterns. Sister cells from the same 8-cell embryo are highlighted together. c) Scatterplot showing the number of genes that are significantly differentially expressed (FDR < 5% and at least 2-fold change between oocytes and 1-cell stage embryos (pronuclei and zygotes). d) Heatmap showing expression patterns of 79 genes activated in the one cell stage throughout all stages of preimplantation.

**Fig. 2. Tracing parent-of-origin allelic RNA transcript via single nucleotide variant analysis in preimplantation embryos.**

a) Schematic for deducing maternal/paternal transcript origin from presence of absence of homozygous paternal allele. In the case of embryo #3, it is uncertain whether the maternal allele also carries this SNP pattern. ii) In the case of known imprinted transcripts, paternal transcripts can be identified if the gene has been previously classified as paternally expressed.

b) Pie charts showing the number and percent of all gene transcripts exhibiting SNVs and their assignments into one of three categories as illustrated in Fig. 2a: 1) mono-allelic maternally expressed pattern (red), 2) biallelic (blue), and 3) uncertain (green).

c) Example of a deduced partial haplotype. Paternal activation was determined by comparing embryos that carried bi-allelic expression in one stage but mono-allelic (maternal) in a previous stage.
d) Example of paternal activation by examining the presence or absence an allele that is paternally homozygous as determined by Exome-seq. rs3829009 is highlight because the minor “T” allele is an Arg884Ser missense variant (see Supplementary Text).

**Fig. 3. Network analysis of human pre-implantation development.** a) Hierarchical cluster tree showing co-expression modules identified using WGCNA. Modules correspond to branches and are labeled by colours as indicated by the first colour band underneath the tree. The remaining colour bands reveal correlations between transcripts and stage indicators. Red and green indicates that the transcript is highly correlated or anti-correlated for the particular stage, respectively.

b) Heatmap of the modules across samples showing correlation of the 7,313 stage specific genes. Because each developmental window only has 1 or 2 highly correlated modules, the modules were assigned biological names, e.g. for example the bisque4 module is referred to as human oocyte module, the thistle1 module is referred to as human 1 cell module, etc.

c) Boxplots showing the distribution of module expression (mean RPKM of all genes within a given module) for different cell types.

**Fig. 4. Stage-specific gene activation is preserved in human and mouse preimplantation development.** a) Heatmap of preservation scores between human stage-specific modules in independent human and mouse datasets found in the literature (x-axis). Module colours are as defined in Supplementary Fig. 2.

b) Heatmap showing the gene overlaps significance between independently constructed human and mouse modules. Modules that are stage-specific have been labeled with the biologically
relevant name, and colors are as defined in Supplementary Fig. 2 and 3. X-axis shows only human stage-specific genes and y-axis shows all mouse modules. Each cell contains the number of intersecting genes and p-value of the intersecting gene lists. Legend represents –log10 p-value based on the hypergeometric test.

c) Schematic drawing illustrating the sequential transcriptome switch of four major pathways (cell cycle, gene transcription, RNA processing and translation, and mitochondrial function) in human and mouse preimplantation embryos. Gene lists for gene ontology terms are derived from human and mouse stage-specific modules with significant overlap ($p<10^{-4}$) as shown in Fig. 4b.

d) Module visualization of hub genes and associated function. ViSANT was used to visualize network connections of the top 150 connections in the darkslateblue (4-cell) module. Highly connected intramodular hub genes (connected to at least 25% of the network) are coloured in red. Independently validated hub genes (kME > 0.9, $p<10^{-22}$) are highlighted in blue (for human only), and orange (for both human and mouse datasets).
Methods

Ethics Statement

Oocytes were obtained from female patients (between ages 26 to 35) at the Center for Clinical Reproductive Medicine at the Jiangsu People’s Hospital/First Affiliated Hospital of Nanjing Medical University through written informed consent and institutional approval. Sperm was obtained from an anonymous healthy donor at a sperm bank in Changsha, China with informed consent. This study was approved by the Institutional Review Board (IRB) on Human Subject Research and Ethics Committee in the First Affiliated Hospital to Nanjing Medical University and Tongji University in China. None of the donors received any financial re-imbursement. Patient IDs were anonymous to all the scientists in this project from the very beginning. Investigators at UCLA were only involved in data analysis and manuscript writing. According to UCLA IRB review of their involvement in the study, UCLA investigators were granted exemption approval (IRB#12-001361).

Human embryo collection and culture

Vitrified mature oocytes were warmed with a thawing kit before RNA isolation (Jieying Laboratory). All thawing steps were performed at room temperature. Oocyte survival was evaluated based on integrity of the oocyte membrane and the zona pellucida. Fertilization is achieved through intracytoplasmic sperm injection (ICSI) in order to synchronize the development time for each batch of viable oocytes for this study. Embryos were cultured in Cleavage Medium (SAGE, USA) in a low-oxygen humidified atmosphere containing 5% CO₂, 5% O₂ and 90% N₂ at 37°C. Stages of human embryo development were assessed through microscopy and collected separately.

Mice embryos collection
PMSG (7-10UI) was intraperitoneally injected into 4-6 weeks C57BL/6 female mice. hCG (7-10UI) was then intraperitoneally injected after 46-48h. Pregnant mice were sacrificed at various time points to obtain embryos as follows: metaphase II oocyte (12h), zygote (24-26h), 2-cell (30-32h), 4-cell embryo (Day 2), 8-cell embryo (Day 3), and morula (Day 4).

**RNA isolation and library construction**

Both human and mouse blastomeres were prepared using identical protocols. Single blastomeres were isolated by removing the zona pellucida using acidic tyrode solution (Sigma, cat. no. T1788), then separated by gentle mouth pipetting in a calcium-free medium. Single cells were washed twice with 1x PBS containing 0.1% BSA before placing in lysis buffer. RNA was isolated from single cells or single morula embryos and amplified as previously described. Library construction was performed following Illumina manufacturer suggestions. Libraries were sequenced on the Illumina Hiseq2000 platform and sequencing reads that contained polyA, low quality, and adapters were pre-filtered prior to mapping. Filtered reads were mapped to the hg19 genome and mm9 genome using default parameters from BWA aligner, and reads that failed to map to the genome were re-mapped to their respective mRNA sequences to capture reads that span exons.

**Transcriptional profiling**

In both human and mouse cases, data normalized was performed by transforming uniquely mapped transcript reads to RPKM. Genes with low expression in all stages (average RPKM<0.5) were filtered out, followed by quantile normalization. For differential expression, we compared every time point to its previous time point using default parameters in DESeq using normalized read counts. Genes were called differentially expressed if they exhibited a Benjamini and Hochberg-adjusted p-value (FDR) <5% and mean fold-change > 2.
**Paternal exon sequencing**

The paternal genomic DNA was extracted from 10ml peripheral blood of the sperm donor (serial number: D5422) using QIAamp DNA Blood Mini Kit. Then 1ug of the genomic DNA was fragmented via sonication. Exome enrichment was performed using the Agilent SureSelect Human All Exon Kit (50Mb). Sequencing was performed on the Illumina Hiseq2000 platform per manufactures’ instructions and reads were mapped to the hg19 using default parameters for the BWA aligner\(^{25}\).

**SNP analysis**

SNP calling was performed using the GATK software using default parameters\(^{16}\). The SNP database we used for reference was from the 1000 genomes project which contains SNP variants from the Chinese population (CHB). Only SNP sites with coverage of 10 reads or more were considered. SNP annotation was performed used Annovar\(^{27}\) and SNP variant effect on protein functional change was predicted using SIFT\(^{28}\). Only predictions with high confidence (SIFT score < 0.05) were considered.

**Paternal and Maternal gene calling**

We performed paternal and maternal gene calling in two ways. In the first, we deduced the parental haplotype and used linkage information to determine parent-of-origin. We used heterozygous SNPs from paternal genotype (first allele>=5 reads, second allele>=3 reads, both quality values >20) and identified at least two embryos which contain paternal SNPs (both alleles>=5 reads, both quality values >20). Next, we inferred father's haplotype using paternal alleles from embryos. We asked linkage between two adjacent loci should be supported by correct links in at least two embryos, and the correct links should be significant more than the
wrong links (such as correct links > 3*wrong links). We scanned all loci and defined some linkage blocks. These results are in Supplemental Table 2.

In the second approach, we used homozygous SNPs from the paternal genotype so we can trace exactly which allele the embryo must carry. If embryos are heterozygous at this site, we infer the alternative allele is maternally derived. In addition, if there is absence of the paternal allele and expression of an alternative allele, these transcripts are also maternally derived. We have provided a schematic for these scenarios in Fig. 3a.

**Weighted Gene Co-expression Network Analysis (WGCNA)**

Both human and mouse datasets were independently constructed using the following method. A signed weighted correlation network was constructed by first creating a matrix of pairwise correlations between all pairs of genes across the measured samples. Next, the adjacency matrix was constructed by raising the co-expression measure, 0.5 + 0.5*cor, to the power $\beta = 12$. The power of 12, which turns out to be the default value, is interpreted as a soft-threshold of the correlation matrix. Based on the resulting adjacency matrix, we calculated the Topological Overlap, which is a robust and biologically meaningful measure of network interconnectedness (i.e. the strength of two genes’ coexpression relationship with respect to all other genes in the network). Genes with highly similar coexpression relationships were grouped together by performing average linkage hierarchical clustering on the topological overlap. We used the Dynamic Hybrid Tree Cut algorithm to cut the hierarchal clustering tree, and defined modules as branches from the tree cutting. We summarized the expression profile of each module by representing it as the first principal component (referred to as module eigengene). Modules whose eigengenes were highly correlated (correlation above 0.7) were merged.

**Module preservation statistics**
To evaluate the human modules in mouse developmental data\textsuperscript{20}, we mapped human genes to mouse genes (orthologous genes) as annotated from the Mouse Genome Informatics (MGI) database\textsuperscript{31}. An advantage of WGCNA is that it provides powerful module preservation statistics which assess whether the density (how tight interconnections among genes in a module are) and connectivity patterns of individual modules (e.g. intramodular hub gene status) are preserved between two data sets. To assess the preservation our human modules (reference network) in the test network (either human or mouse), we used the R function modulePreservation() in the WGCNA R package since this quantitative measure of module preservation allows one to rigorously argue that a module is not preserved\textsuperscript{20}. By averaging the several preservation statistics generated through many permutations of the original data, a $Z_{\text{summary}}$ is calculated, which summarizes the evidence that a module is preserved and indicative of module robustness and reproducibility. In general, modules with $Z_{\text{summary}}$ scores $> 10$ are interpreted as strongly preserved (i.e. densely connected, distinct, and reproducible modules), $Z_{\text{summary}}$ scores between 2 and 10 are weak to moderately preserved, and $Z_{\text{summary}}$ scores $< 2$ are not preserved.

**Identification and visualization of hub genes**

Module eigengenes lead to a natural measure of module membership (also known as module eigengene based connectivity kME). Specifically, a fuzzy measure of module membership for gene $i$ with respect to module $q$ is defined as follows $\text{MM}^q(i) = \text{cor}(x(i), E^q)$, where $x(i)$ is the expression profile of gene $i$ and $E^q$ is the eigengene of module $q$. Since we use signed networks here, we expect that module genes have significant positive MM values. The advantage of using a correlation to quantify module membership is that this measure is naturally scaled to lie in the interval [-1, 1] and a corresponding statistical significance measure (p-value) can be easily computed. Genes with highest module membership values are referred to as intramodular hub
genes (e.g. kME > 0.9, \( p < 10^{-22} \)). Intramodular hub genes are centrally located inside the module and represent the expression profiles of the entire module\(^\text{32}\). We used VisANT\(^\text{33}\) to visualize the top 150 gene connections (based on topological overlap) among the top 100 hub genes (i.e. genes with highest kME).

**Hub gene validation**

We used WGCNA to independently construct a network in published datasets, and generated an independent list of hub genes (kME > 0.9) for each dataset. For human-human comparisons and mouse-mouse comparisons, we determined the overlap of hub genes from the same developmental stage. However, for human-mouse module overlaps, we used all modules with significant gene overlap (\( p < 10^{-4} \), see Fig. 4b) to compute intramodular hub gene overlap. For example, we overlap hub genes found in the human darkslateblue module to both mouse oocyte and 1-cell networks since both these networks have significant overlaps with darkslateblue.

**Enrichment of upstream regulators**

Ingenuity was used to determine enrichment of upstream regulators for all human stage-specific modules. Regulators which were not transcriptional regulators were not considered.
References


22 Rother, F. et al. Importin alpha7 is essential for zygotic genome activation and early mouse development. *PloS one* 6, e18310.

23 Dannenberg, J. H. et al. mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival. *Genes & development* 19, 1581-1595


Figure 8-1.
Figure 8-2.

(a) Phased paternal SNP pattern

<table>
<thead>
<tr>
<th>Origin</th>
<th>Maternal</th>
<th>Paternal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

(b) Known imprinting gene

<table>
<thead>
<tr>
<th>Paternal genotype</th>
<th>8-cell</th>
<th>morula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>115</td>
<td>216</td>
</tr>
<tr>
<td>No</td>
<td>753</td>
<td>614</td>
</tr>
</tbody>
</table>

(c) Paternal haplotype

<table>
<thead>
<tr>
<th>ASB6</th>
<th>PRX2</th>
<th>PTGES</th>
<th>TOR1B</th>
<th>TOR1A</th>
<th>C1orf78</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
</tbody>
</table>

(d) CDC28 exon XIII

<table>
<thead>
<tr>
<th>Paternal genotype</th>
<th>rs38209009</th>
<th>rs48890278</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6537</td>
<td>rs48890278</td>
<td>rs6537</td>
</tr>
</tbody>
</table>

Transcript RPKM
Figure 8-3.

(a) A graph showing data points on the y-axis labeled "height" and the x-axis labeled "colors." The graph includes data points for H. oocyte, H. 1-cell, H. 2-cell, H. 4-cell, and H. 8-cell.

(b) A heatmap showing gene expression levels across different stages of development: oocyte, 1 cell, 2 cell, 4 cell, and 8 cell. The heatmap includes color-coded categories for various biological processes.

(c) A series of bar charts showing module RPKM (Reads Per Kilobase Million) for different stages: oocyte, 1 cell, 2 cell, 4 cell, and 8 cell. The charts are color-coded for oocyte (bisque4), 4-cell (darkslateblue), and 8-cell (darkslateblue).
Figure 8-4.

(a) Human Module Preservation

(b) Human-Mouse Module Overlap

(c) Molecular networks in human and mouse embryos

(d) Gene expression networks in human and mouse embryos
DNMT3B mutations result in selective demethylation and altered gene expression associated with ICF Syndrome

Kevin Huang¹, Zhenshan Liu², Zhourui Wu¹,², Juehua Yu¹, Zhigang Xue¹,², Thuc Le¹, Kee-Pyo Kim¹, Nagesh Rao⁴, Matteo Pellegrini³, Guoping Fan¹*

¹Department of Human Genetics, Broad Stem Cell Research Center, David Geffen School of Medicine, University of California Los Angeles, 695 Charles Young Drive, Los Angeles CA, 90095 USA; ²Translational Stem Cell Center, Tongji Hospital and Department of Regenerative Medicine, Tongji University School of Medicine, Shanghai, China; ³Department of Molecular, Cell and Developmental Biology and ⁴Department of Pathology, University of California Los Angeles, California 90095 USA
Abstract

High resolution DNA methylation maps of somatic and pluripotent cells have revealed many changes in the methylation profile that occur during cellular reprogramming, including a wave of de novo methylation. However, it is unclear how the de novo DNA methyltransferases contribute to shaping the pluripotent methylome during the process. Here, we generated iPSCs carrying double heterozygous DNMT3B point mutations from ICF Syndrome patient fibroblasts and performed whole genome bisulfite sequencing. Because DNMT3A and DNMT3B are both highly expressed in normal human iPSCs, we determined DNMT3A can functionally compensate for most CG sites in DNMT3B-deficient ICF iPSCs. However, we identified large megabase domains typically localized near centromeres and subtelomeres that were selectively demethylated, indicating these are unique DNMT3B targets. Furthermore, we detected severe global loss of non-CG methylation, suggesting DNMT3B is the primary enzyme responsible for establishing non-CG methylation in reprogrammed cells. Also as a consequence of DNMT3B deficiency, ICF-iPSCs exhibited abnormal gene expression relevant to the ICF Syndrome phenotypes. Many of the up-regulated genes were associated with selective and localized promoter demethylation in ICF iPSCs, suggesting these genes are exclusively repressed by DNMT3B. Collectively, our results uncovered a crucial role for DNMT3B in regulating DNA methylation during cellular reprogramming, and established a valuable system to study ICF Syndrome pathogenesis in vitro.
Introduction

DNA cytosine methylation is one of the major epigenetic factors that regulates many developmental events including genomic imprinting, X-inactivation, genome defense, and developmental gene regulation (Jaenisch and Bird, 2003). DNA methylation is established and maintained by a family of DNA methyltransferases (DNMTs) including DNMT1, DNMT3A and DNMT3B (Goll and Bestor, 2005). Aberrant methylation patterns have been associated with a number of human diseases such as cancer, immunodeficiency, and neurological disorders (Feng and Fan, 2009). For example, DNMT1 mutations are linked to hereditary sensory and autonomic neuropathy type 1 (HSAN1) disease in which patients develop dementia and hearing loss in adulthood (Klein et al., 2011). DNMT3A mutations are coupled with subtypes of leukemia (Ley et al., 2010; Shah and Licht, 2011), and DNMT3B mutations cause a rare genetic disorder namely ICF Syndrome (Immunodeficiency, Centromere instability, and Facial anomalies) (Okano et al., 1999; Xu et al., 1999). Relevant to the loss of DNMT3B activity, lymphoblasts from ICF patients exhibit hypomethylation of pericentromeric regions and selective demethylation of promoters associated with altered gene expression (Ehrlich et al., 2008; Jin et al., 2008).

DNA methylation also plays an important role in stem cell self-renewal and differentiation (Bibikova et al., 2008; Huang and Fan, 2010). We and others have demonstrated that DNA methylation contributes to tissue- and cell-specific gene expression patterns in parallel with other known regulatory mechanisms such as transcriptional networks and histone modifications (Ernst et al., 2011; Fouse et al., 2008; Hawkins et al., 2010; Meissner et al., 2008; Young, 2011). Upon differentiation of both mouse and human embryonic stem cells (mESCs and hESCs), a wave of de novo methylation takes place that silences pluripotency genes and
establishes tissue-specific methylation patterns (Doi et al., 2009; Hawkins et al., 2010; Lister et al., 2009; Meissner et al., 2008; Mohn et al., 2008; Pai et al., 2011; Shen et al., 2006; Stadler et al., 2011). In the absence of Dnmts, demethylated mESCs undergo apoptosis soon after inducing in vitro differentiation (Farthing et al., 2008; Tsumura et al., 2006). More recently, there has been an increasing interest in the function of DNA methylation during cell reprogramming. It has been postulated that epigenetic barriers such as DNA methylation have to be overcome in order to achieve successful reprogramming of induced pluripotent stem cells (iPSCs). Indeed, selective promoter demethylation of pluripotent genes such as Oct4 and Nanog are associated with successful reprogramming of somatic cells to iPSCs (Bhutani et al., 2010; Chan et al., 2009; Huangfu et al., 2008; Mikkelsen et al., 2008; Wernig et al., 2008). In addition, inhibiting DNMTs activities with 5-aza-cytidine (AzaC), or DNMT1 knock-down promotes partially reprogrammed cells into fully reprogrammed iPSCs (Huangfu et al., 2008; Mikkelsen et al., 2008; Shi et al., 2008). On the other hand, a wave of de novo methylation also occurs during reprogramming. Tissue-specific genes and so-called partially methylated domains becomes hypermethylated (Doi et al., 2009; Lister et al., 2011), and hallmark enrichment of non-CG methylation becomes re-established in iPSCs (Lister et al., 2009; Lister et al., 2011). Together, these results indicate major methylome alterations underlie cellular reprogramming.

To understand how DNMT3B contributes to the methylation changes during cellular reprogramming, we generated iPSCs from human ICF Syndrome patient fibroblasts carrying DNMT3B mutations and profiled methylation patterns through genome-scale bisulfite sequencing. We identified several unique targets of DNMT3B at both the large-scale megabase domains as well as selective promoter demethylation. ICF iPSCs also exhibit dramatic loss of non-CG methylation, indicating DNMT3B is the major enzymes for catalyzing non-CG
methylation. Finally, through RNA-seq we identified altered gene expression in ICF iPSCs linked to the disease phenotype in ICF Syndrome patients, suggesting that ICF iPSCs would be a valuable tool to study ICF pathogenesis in vitro.

Results

DNMT3B is not required for direct reprogramming of human fibroblasts

Induced pluripotent stem cells exhibit higher levels of methylation compared to somatic cells, suggesting reprogramming involves a wave of de novo methylation. Because Dnmt3B is more dramatically up-regulated in iPSCs when compared to levels of Dnmt1 and Dnmt3A (Meissner et al., 2008), we hypothesized DNMT3B may play a major role for de novo methylation in human iPSCs. To directly examine this hypothesis, we generated iPSCs derived from two lines of Immunodeficiency, Centromere instability, and Facial anomalies (ICF) syndrome patient skin fibroblasts carrying double heterozygous point mutations in the catalytic domain of DNMT3B (Figure 9-1A). Similar to Dnmt3b deficient mouse somatic cells, ICF fibroblasts can also be reprogrammed with no obvious difference in reprogramming efficiency (Pawlak and Jaenisch, 2011) (Figure 9-1B). Immunostaining and RT-PCR demonstrated ICF iPSCs expressed typical markers for pluripotency equivalent to normal iPSCs and hESCs (Figure 9-1B and Figure S1). Significantly, ICF iPSCs form teratomas in vivo, demonstrating these reprogrammed cells can be differentiated into all three germ layer lineages (Figure 9-1C). Interestingly, in contrast to major chromosomal instabilities seen in ICF patient B-cells, ICF iPSCs karyotype appeared normal (Figure 9-1D). However, chromosomal epigenetic aberrations such as hallmark hypomethylation at Satellite 2 repeats were consistently demethylated in ICF iPSCS (Figure 9-1E). Overall, we
conclude that although DNMT3B is dispensable for reprogramming, but retains somatic methylation alterations.

**DNMT3B contributes to de novo methylation during cellular reprogramming**

HPLC-MS analysis revealed ICF iPSCs exhibit significantly reduced levels of global methylation when compared to normal iPSCs, but still higher levels when compared to somatic cells (Figure 9-2A). This data would suggest DNMT3A also contributes to some de novo methylation during cell reprogramming. To map the DNA methylation alterations in ICF iPSCs at base-pair resolution, we performed whole genome shotgun bisulfite sequencing in two separate ICF patient iPSCs and achieved an average coverage of 10X (ICF1-1 iPSC) and 4X (ICF1-2 iPSC) per strand. Globally, our whole genome bisulfite sequencing estimated similar loss of global methylation in both ICF iPSC lines (Figure 9-2B). Remarkably, our base-pair resolution revealed non-CG methylation levels were drastically reduced whereas CG methylation was only mildly diminished (Figure 9-2B).

Compared to normal hiPSCs, we found nearly complete loss of methylation in both CHG and CHH motifs but less than 10% loss in CG motifs. Loss of non-CG methylation appeared to be uniform across chromosomes and genes (Figure S2), and can also be confirmed using conventional bisulfite sequencing (Figure S3). Although the vast majority of non-CG methylation was abrogated in ICF iPSCs, we did detect some residual non-CG sites that retained high levels of methylation. Indeed, compared to their parental fibroblast cell lines, both ICF1-1 and ICF1-2 iPSCs had detectably greater levels of non-CG methylation (Figure 9-2C). Strikingly, the few remaining non-CG sites that were highly methylated were significantly conserved in both ICF iPSCs, suggesting DNMT3A has a minor role for targeting non-CG sites.
(Figure 9-2D). Sequence motif analysis of the highly methylated non-CGs in control iPSCs reveals an enriched TACAG motif, consistent with previous reports (ref) (Figure 9-2D). However, the TACAG motif is abolished in the highly methylated non-CG sites in ICF iPSCs, but a strong CA motif is remarkably still present. Collectively, these data indicate that whereas DNMT3A can redundantly compensate for DNMT3B-deficiency at most CpG sites, DNMT3B is the foremost enzyme in establishing non-CG methylation in human iPSCs.

**DNMT3B targets large megabase domains during cellular reprogramming**

Our global methylome analysis revealed obvious large-scale regions of aberrant CG methylation that were consistent in both ICF iPSCs, suggesting these two disease iPSCs share deficiency in similar pathways (Figure 9-3A). In total, we identified 110Mb demethylated megabase regions that showed at least 30% reduced CG methylation (Table S1). These megabase domains tend to undergo de novo methylation during methylation since somatic cells tend to be relatively hypomethylated compared with normal iPSCs (Figure 9-3B). Interestingly, these regions were commonly located at centromeric and subtelomeric regions which are generally regarded as transcriptionally inactive for mRNA (Figure 9-3A). As expected, Satellite 2 repeats found enriched at pericentromeric regions were confirmed to be selectively demethylated compared to other repeat motifs (Figure S4). Of note, 85% of these total megabase regions were also hypomethylated in ICF lymphoblasts compared to control lymphoblasts (Figure S5), indicating these regions are also established by DNMT3B in somatic cells. Most strikingly, we observed massive and nearly uniform 50% demethylation across the X-chromosome (Figure S6). Surprisingly, although we did not observe any obvious changes in total X-chromosome RNA abundance, our gene expression analysis revealed that many up-regulated genes are skewed
toward the X-chromosome. As we will discuss later, this result indicate that these female ICF iPSCs still has an inactive X chromosome.

We were next curious how DNMT3B deficiency affected other megabase iPSC-specific DNA methylation features reported in the literature. Recently, “hotspots” for aberrant methylation have been confirmed in multiple iPSCs, namely 11 and 121 regions of hyper- and hypomethylation in iPSCs (Lister et al., 2011). Strikingly, we found that all hypermethylated DMRs in normal iPSCs were hypomethylated in mutant ICF iPSCs (Figure 9-3C), whereas hypomethylated DMRs in ICF cells remained hypomethylated, reinforcing the notion that DNMT3B is primarily associated with de novo methylation in iPSCs. Next, we examined the methylation status at large scale regions of reduced CpG methylation in somatic cells, termed partially methylated domains (PMDs), which normally become fully methylated in iPSCs (Lister et al., 2011). However, PMDs in ICF iPSCs showed a range of methylation levels, indicating some PMDs do not fully re-acquire DNA methylation in the absence of DNMT3B whereas others can be compensated by DNMT3A in ICF iPSCs (Figure 9-3D). We also looked into nuclear lamina associated domains (or LADs). Although LADs methylations levels are comparable between normal and ICF iPSCs, we found dramatic LAD hypomethylation in ICF lymphoblasts, indicating DNMT3B play a crucial role in methylating these regions in the hematopoietic lineage (Figure S6).

Remarkably, we also found that our megabase regions of CpG demethylation in ICF-iPSCs had a significant overlap (~80%) with previously identified non-CG DMRs (Lister et al., 2011) (Figure 9-3E-F). In normal iPSCs, non-CG DMRs are megabase domains showing unaltered levels of CG methylation, but often lower levels of non-CG methylation when compared to ESCs. Importantly, somatic cells also show low levels of CpG methylation in these
genomic regions (Lister et al., 2011). Therefore, the partial reduction of CpG methylation in ICF-iPSCs supports the notion that DNMT3B is involved in de novo CpG methylation at non-CG DMRs in normal iPSCs (Figure 9-3E-F). We conclude that DNMT3B preferentially targets CpG methylation instead of non-CG methylation in these megabase domains that uniquely exhibit lower levels of non-CG methylation when compared to other genomic regions.

**DNMT3B-deficient iPSCs exhibit deregulated gene expression relevant to ICF Syndrome**

To determine the effects of DNMT3B-deficiency on gene expression in iPSCs, we performed mRNA-Seq on ICF iPSCs using both hESCs and hiPSCs as controls. To carefully identify the gene expression change in ICF iPSCs, we performed pair-wise comparisons between ICF iPSCs and control hESCs and hiPSCs. Using criterion of FDR < 1% and > 1.5 fold, we identified 136 and 213 up- and down-regulated genes that were consistently deregulated in both ICF2 and ICF3 iPSCs (Figure 9-4A, Table S1). Although gene ontology analysis revealed no functional enrichment for the set of upregulated genes, the downregulated set of genes were significantly enriched for terms in cell adhesion (Figure S7).

Methylation profile of these altered genes showed highly selective promoter demethylation, but not gene body demethylation (Figure 9-4B). Interestingly, 36 (26%) upregulated genes showed at least 20% demethylation of the proximal promoter, indicating a potential direct role for DNMT3B (Figure 9-4C). Downregulated genes also showed some promoter demethylation, but this is due to massive unselective demethylation of the entire protocadherin cluster (Figure 9-4C and Figure S8). We conclude these protocadherin genes may be downregulated by some indirect mechanism.
We next examined the genomic context of these altered gene expression and found that the majority of altered gene promoters are biased for CGI promoters (Figure 9-4D). For example, gene promoters such \textit{CHCHD2}, \textit{ZNF248}, \textit{ZNF560}, showed selective demethylation at the CpG island domain in the proximal promoter (Figure S9-11). By contrast, gene promoters such as \textit{CXorf61} and \textit{MAGEE2} have low CG density (Figure S12-13). Therefore, DNMT3B seem to be uniquely targeted to repress sets of genes irrespective of promoter CG density. In addition, we looked at other chromatin modifiers that may explain the altered gene expression (Figure 9-4D).

In our analysis, promoters at up-regulated genes do not show any enriched chromatin marks. By contrast, we observed that 91% of down-regulated genes were highly enriched for H3K27me3 and H3K4me1 (Figure 9-4D). H3K4me1 is a common marker for enhancer sites in conjunction with H3K27ac, and methylation analysis revealed ICF iPSCs showed selective demethylation at enhancer sites (Figure S14).

Gene expression profiles have been previously studied in ICF patient lymphoblastoid cell lines (Ehrlich et al., 2001; Jin et al., 2008). These studies reported genes involved in immune function and neurogenesis were deregulated (Ehrlich et al., 2001; Jin et al., 2008). To determine whether deregulated genes in ICF lymphoblastoids can be seen earlier in development, we cross-referenced our set of deregulated genes in ICF iPSCs with previous studies. Commonly upregulated genes between ICF iPSCs and ICF lymphoblastoids included \textit{ANXA3}, \textit{CASK}, \textit{GLT1D1}, and \textit{TSPYL5}. In ICF iPSCs, both GLTD1 and TSPYL5 showed selective promoter demethylation, indicating these two targets may be direct targets of DNMT3B (Figure S15).

Commonly down-regulated genes include \textit{GPR160}, \textit{JAKMIPI}, \textit{S100A11}, and \textit{HCK}. \textit{HCK} is a member of the Src-family of tyrosine kinases primarily involved in the hematopoietic lineages, consistent with a possible defect in the hematopoietic signaling cascade in ICF-iPSCs. In
addition, we also found the class of protocadherin genes was significantly down-regulated (Figure S8), indicating deregulation of the protocadherin (PCDH) cluster began in the stem cell stage and persisted in differentiated cells as observed in lymphoblastoids. Deregulation of the PCDH gene cluster may be associated with the neural deficiency observed in ICF Syndrome patients. Consistently, we found two glutamate receptors GRIA1 and GRIN1 both downregulated in ICF iPSCs. Together, our results demonstrated ICF-iPSCs exhibit characteristic features of gene expression associated with disease phenotypes of ICF syndrome.

**Discussion**

Through reprogramming ICF patient fibroblasts carrying DNMT3B mutations, we discovered that DNMT3B is necessary for many *de novo* methylation signatures, but is dispensable for overall iPSC derivation. Our finding reinforces the findings from a recent study that showed mouse iPSCs can be readily generated from embryonic fibroblasts deficient of either Dnmt3a, or Dnmt3b, or both (Pawlak and Jaenisch, 2011). Thus, although a wave of *de novo* methylation in cell reprogramming contributes to iPSC-specific methylation, demethylation in pluripotency genes is more associated with pluripotency. Consistently, mouse ESCs can tolerate total demethylation while maintaining stem cell self-renewal (Tsumura et al., 2006; Hutnick et al., 2010).

In hESCs, Ziller et al. (2011) demonstrated a strong correlation between non-CpG methylation and DNMT3 expression levels through siRNA knock-down experiments. In this study, we showed that DNMT3B mutations lead to the loss of over 90% of non-CG methylation, indicating DNMT3B is the predominant *de novo* methyltransferase for non-CG methylation. Of note, we have also performed genome bisulfite sequencing in Dnmt3−/− and Dnmt3b−/− mESCs
(Huang et al., in preparation), and found non-CG methylation is equivalently and moderately reduced in both cell lines, suggesting Dnmt3a and Dnmt3b are functionally redundant for non-CG methylation in mice. This would also suggest the regulation of non-CG methylation is species specific. Of interest, non-CG methylation has previously been positively correlated with expression. However, our RNA-Seq analysis in ICF iPSCs did not reveal any obvious changes to genes that were previously enriched for non-CG methylation. Therefore, it is still unclear as to the function for non-CG methylation in human pluripotent stem cells.

Our gene expression analysis in ICF-iPSCs allowed us to identify several genes that are deregulated both in ICF patient somatic cells and iPSCs. Most notably, loss of cell adhesion appears to be a recurrent theme underlying ICF transcriptome in all cell types. In addition, we found that several genes required for lymphocyte differentiation and genome stability are already deregulated in ICF iPSCs, suggesting that altered lymphocyte cell development in ICF patients can be predicted from altered gene expression in mutant ICF-iPSCs. Thus, the derivation of ICF-iPSCs also provides a valuable system to study ICF Syndrome pathogenesis through in vitro differentiation into a variety of somatic cell types relevant to ICF Syndrome phenotypes such as hematopoietic cells, neural crest lineage cells, and mesenchymal cells.

**Material and Methods**

**Derivation and cultures of human iPSCs and hESCs**

The production of human iPSCs follows up the protocol described by Takahashi et al. (2007) and Yu et al. (2007) using retroviruses expression OCT4, SOX2, KLF4, and c-MYC or OCT4, NANOG, KLF4 and LIN-28. hESC cells were maintained in DME supplemented with 20% KSR, nonessential amino acids (Invitrogen), L-Glutamine (Mediatech), Penn/Strep, 2-
mercaptoethanol with a feeder layer of MEFs as previously described (Shen et al., 2006). For both gene expression and methylation analysis studies, the hESCs were passaged onto feeder free gelatin coated plates twice before harvesting RNA and DNA. RNA was isolated using Trizol (Invitrogen) while DNA was isolated using PureLink™ genomic DNA purification kit (Invitrogen).

**Southern blot analysis**

Genomic DNAs were digested with HpaII enzymes overnight and fractionated through agarose gel electrophoresis and transferred to nitrocellular membrane. Satellite 2 and 3 probes were end-labeled with $^{32}$P radioisotope and were hybridized with Quickhyb (Stratagene, San Diego), following up the protocol of manufacturer. Radioautography was performed with hybridized membrane and developed with a Kodak film developer.

**Bisulfite Conversion and Sequencing**

Bisulfite conversion was performed as described (Shen et al., 2006). Briefly, we digested genomic DNA with BglII overnight. Digested DNAs were then incubated with a sodium bisulfite solution for 16 hours. Bisulfite treated DNA was then desalted and precipitated. We used 1/10 of precipitated DNA for each PCR. For PCR, we used nested primers to generate our products. PCR products were gel purified and used for Topo Cloning (Invitrogen).

RRBS and whole genome shotgun bisulfite sequencing libraries were prepared as previously described (Bock et al., 2011; Lister et al., 2011) and sequenced on the Illumina HiSeq2000 machine following manufacturer instructions.

**Data analysis**

Take raw read counts, normalize by scaling. Quadruple intersect of Poisson cdf.
Quantitative Real-time PCR
RNA was DNase I treated (Invitrogen) and then quantified again. cDNA conversion was done using the iScript kit (BioRad). Quantitative PCR was done on a MyIQ Thermocycler (Biorad) using the Sybr Green Supermix (BioRad). Relative expression levels were normalized to 18s amplicons.

Accession Number:
Microarray and high-throughput sequencing data will be available to the public at GEO (Gene Expression Omnibus) upon publication of this paper.

Acknowledgements
Funding for this work is from NIH R21DE02298. KH is supported by the California Institute of Regenerative Medicine Training Grant (TG2-01169).

Figure legends
Figure 9-1. Generation of DNMT3B-deficient iPSCs.
A. Schematic of the double heterozygous point mutations in two separate ICF Syndrome patients.
B. Immunostaining of pluripotency markers in ICF iPSCs C. Teratoma histology for three different germ layers. D. Karyotype of ICF patient fibroblasts and reprogrammed iPSCs. E. Southern blot analysis probing CpG DNA methylation for Satellite 2 repeats

Figure 9-2. DNMT3B is the dominant methyltransferase for non-CG sites.
A. HPLC-MS analysis of total methylcytosine levels in various cell types. B. Bar chart showing global methylation in CG, CHG, and CHH motifs as measured by whole genome shotgun
bisulfite sequencing. Only sites shared were considered. C. Bar plot showing non-CG methylation in iPSCs and parental fibroblasts. As additional control, BJ1 fibroblasts are used to show non-CG methylation in normal fibroblast cells. All fibroblasts were measured via RRBS. D. Highly methylated non-CG sites (>0.3) are highly conserved with significant sequence motif. Motif logo was generated by via.

**Figure 9-3. DNMT3B-deficient partially ablates iPSC signature methylation profile**

A. Circos plot showing genome-wide megabase view. Color indicates methylation level. B. Heatmap of all CG-DMRs found in ICF iPSCs compared to control iPSCs and parental fibroblasts. C. Heatmap of aberrantly hypermethylated regions in iPSCs as previously identified (n=11, Lister et al., 2011). D. Boxplot showing distribution of methylation levels in 10kb regions within shared PMDs (left) and associated transcript abundance (right). E. Genome browser view of methylation level in a representative normal iPSCs and ICF1-1 iPSCs. Note coincidence of reduced non-CG methylation in IMR90 iPSCs and reduced CG methylation in ICF2 iPSCs. F. Metaplot analysis of CG methylation in non-CG DMRs. ESCs, iPSCs, and somatic cells were averaged over the dataset as shown in Figure 3B.

**Figure 9-4. Altered gene expression is associated with promoter demethylation**

A. Scatterplot between the control pluripotent stem cells and ICF iPSCs. Only genes that were consistently up-regulated in both ICF2 and ICF3 were colored. B. Metaplot analysis of the up- and down-regulated genes. C. Scatterplot showing the expression and promoter methylation changes in altered genes. D. Bar graph of the proportion of gene promoters positive for the labeled attributes.
References


Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S.,
Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-

Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P.J., Park, J., Butler, J.,
intermediate states accompany lineage specificity in the adult hematopoietic compartment.
Molecular cell 44, 17-28.

emerging systematic view. Regenerative medicine 5, 531-544.

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., and
only Oct4 and Sox2. Nature biotechnology 26, 1269-1275.


J., et al. (2008). DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to
altered epigenetic modifications and aberrant expression of genes regulating development,
neurogenesis and immune function. Human molecular genetics 17, 690-709.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I.,


181


Figure 9-1.
Figure 9-2.
Figure 9-3.
Figure 9-4.
Chapter 10
Gene body demethylation and H3K27me3 re-distribution contribute to transcriptome deregulation in mutant embryonic stem cells null of DNA methylation

Kevin Huang¹, Juehua Yu¹, Thuc Le¹, Matteo Pellegrini², and Guoping Fan¹*

¹Departments of Human Genetics,
²Molecular, Cell and Developmental Biology,
University of California, Los Angeles
695 Charles Young Drive South,
Los Angeles, CA 90095 USA
Abstract

DNA methylation is postulated to regulate pluripotency and differentiation programs in embryonic stem cells by facilitating expression of pluripotency genes while silencing differentiation genes. We performed mRNA-Seq in multiple control and mutant mouse embryonic stem cells (ESCs) that are deficient in all three DNA methyltransferases (\textit{Dnmt1}^{-/-};\textit{Dnmt3a}^{-/-};\textit{Dnmt3b}^{-/-}, or TKO) and uncovered many deregulated genes involved in various biological processes including differentiation. Global bisulfite sequencing revealed a portion of these genes are likely triggered by direct loss of DNA methylation at proximal promoters. However, we also found a portion of downregulated genes including some housekeeping genes that are minimally methylated in the promoter region, but contain a high level of gene body methylation in wild-type ES cells, suggesting heavy gene body methylation contributes to enhanced gene expression. Furthermore, ChIP-Seq demonstrated that H3K4me3 distribution was unaffected in TKO ESCs whereas H3K27me3 became depleted at many gene promoters, indicating DNA methylation shares pathways with select histone modifications. H3K27me3 depletion occurred primarily at bivalency domains, and loss of bivalency resulted in up-regulation of many genes involved in differentiation. We further re-introduced individual Dnmt into TKO ESCs and found single Dnmts is not sufficient to restoring H3K27me3. Strikingly, global bisulfite sequencing in TKO Dnmt reconstitution mESCs revealed unique and shared genomic targets including novel \textit{de novo} methylation target of Dnmt1. Overall, our study demonstrated that demethylation at either gene promoter or gene body can directly affect gene expression and also disrupt localization of select histone modifications, leading to altered gene regulatory networks in pluripotent stem cells.
Introduction

Embryonic stem cells (ESCs) are a special population of pluripotent cells derived from the inner cell mass of a blastocyst and have the unique ability to indefinitely self-renew and differentiate into any cell type found in the adult body. These unique features of ESCs are maintained by a transcription regulatory network that includes key pluripotent factors, Oct4, Sox2, Tcf3, and Nanog. During cell differentiation, the pluripotency transcription network becomes silenced and lineage-specific transcription networks become activated.

In addition to transcriptional regulatory programs, epigenetic mechanisms such as DNA methylation and histone modifications are important regulators of ESC-specific gene expression and cell differentiation. For example, a number of repressed differentiation genes in ESCs are “bivalently” marked by both active and repressive histone modifications [e.g. trimethylation of H3 lysine 4 and 27 (H3K4me3 and H3K27me3)], and are thought to be “poised” to become activated upon differentiation. Other epigenetic mechanisms such as microRNAs have also been shown to be critical for ESC identity and cell differentiation. It is likely that multiple epigenetic regulators work in concert to fine-tune gene expression, perhaps through both direct and indirect mechanisms.

DNA methylation is one of the best studied epigenetic mechanisms and classically associated with repressing genes, repetitive elements, and parent-specific alleles. In mammals, a family of DNA methyltransferases, including Dnmt-1, -3a and -3b, catalyzes DNA methylation and plays key roles in establishing and maintaining methylation patterns across the genome. Recent genome-wide studies indicate DNA methylation may possess additional properties. For example, in addition to promoter methylation, gene body methylation has emerged as a
regulator of either RNA transcription elongation or a positive factor by antagonizing repressive
effects of polycomb proteins and H3K27me3\(^\text{5,11,12}\).

Using microarray technology, we and others have found that mouse ESCs deficient in
DNA methylation exhibits hundreds of genes up- and down-regulated\(^\text{8}\). We proposed that DNA
methylation acts as a parallel mechanism together with transcription factors, histone
modifications, and other regulatory mechanisms to compose the ESC transcription landscape.
However, we have not addressed the question whether DNA methylation affects histone
modifications and how DNA methylation directly or indirectly influences gene transcription.

Here, we profiled global RNA expression, DNA methylation patterns, and histone
localization patterns in a more comprehensive manner using high throughput sequencing
technology. We report that many more genes are deregulated in methylation deficient ES cells.
Besides the direct effect of DNA methylation in gene silencing and activation, we provide
evidence that indirect regulatory pathways such as histone re-distribution and non-coding RNAs
can indirectly change the ES transcriptome in methylation-deficient ESCs. We also provide
global bisulfite sequencing data for TKO Dnmt reconstitution cell lines and determined the
genome-wide methylation contributions from each DNA methyltransferase.

**Results**

**Global deregulation of mRNA expression in TKO ESCs**

To examine the effects of DNA methylation on global RNA expression in mouse ESCs, we
generated ESCs deficient in all three major DNA methyltransferases (\(Dnmt1^{+/+};Dnmt3a^{+/+}\);
\(Dnmt3b^{+/+}\), or TKO). We have previously demonstrated that TKO cells are virtually devoid of
both DNA methylation and hydroxymethylation\(^\text{13}\). To quantify the consequence of genome-wide
DNA methylation deficiency on RNA transcription, we performed mRNA-seq analysis in two wild-type (WT) and TKO mESCs pairs using either single-end or paired-end sequencing strategies (Table S1). To further enhance our statistical power, we incorporated an additional WT and TKO mRNA-Seq pair from a recently published study. To our surprise, all three independently generated datasets better correlated within their own pairs despite identical genetic background in the cell lines. These patterns highlight variations in datasets that may arise from technical differences including culture conditions, library preparation, and/or laboratory practices. In fact, these non-biological variations (or batch effects) were also previously observed in microarray platforms. We therefore corrected for batch effects to make all WT and TKO cell lines more comparable for downstream analysis (Fig. S1).

Using the criterion of <1% FDR, we found 965 and 1048 genes that are on average at least 1.5 fold up- and down-regulated in TKO ESCs, respectively (Fig. 10-1a). Although a previous analysis of RNA-seq and microarray dataset found a low overlapping (~9%) number of de-regulated genes, we found our dataset has ~40% overlap with previous studies (Fig. 10-1b). Because of the higher concordance with previous datasets, our dataset of gene deregulation is likely the most representative one for TKO cell line to date. We further validated these genes with ~85% accuracy using qRT-PCR in multiple TKO cell lines, and were able to confirm many differentially expressed genes that were previously considered non-significant (Fig. 10-1b, Table S2). Overall, we found many more genes that are differentially expressed than previously thought, indicating DNA methylation has a larger contribution regulating the ES transcriptome.

Consistent with previous studies, we found upregulated genes were enriched in various developmental processes (Fig. 10-1c). For example, the germline specific Rhox family of genes is heavily methylated and silenced in ESCs but becomes demethylated and highly expressed
upon germline differentiation. This region is mostly devoid of any histone marks, suggesting induction of the Rhox cluster in TKO cells was due to direct de-repression (Fig. 10-1d). Other differentiation pathways such as neurogenesis and organogenesis also became upregulated in TKO ESCs, indicating wide-spread activation of tissue-specific genes.

In addition, our dataset was able to detect non-coding RNAs such as lincRNAs, which have recently been shown to be major players in maintaining pluripotency and suppressing differentiation. Many lincRNAs in ES cells become downregulated upon retinoic-acid induced differentiation. Similarly, we found most deregulated lincRNAs are downregulated in the absence of methylation (Fig. S2), supporting the notion that TKO cells are in a state of induced differentiation. Previously, some lincRNAs in ESCs have been shown to be directly regulated by transcription factors Oct4, Sox2, and Nanog (17). However, TKO ESCs have express these transcription factors at nearly the same level as WT ESCs. Therefore, it is still unclear whether these lincRNAs contribute to activation of differentiation genes or indirect targets of DNA methylation.

**DNA methylation profiling of deregulated genes**

To determine the direct relation between RNA expression and DNA methylation, we generated a genome-wide DNA methylation profile of J1 wild-type (WT) ESCs using reduced representation bisulfite sequencing. Our methylation map covers ~1.6 million high confident CGs and ~60 million non-CGs with an average coverage of 30X. By separating wild-type ES expression levels into quantiles, we found a strong inverse correlation between promotor methylation and gene expression as previously reported (Fig. 10-2a). Intriguingly, we also observed a modest and positive correlation between gene body methylation and gene expression, which have previously
been described in other species\textsuperscript{5,12,19}. High expressed genes tend to have higher levels of gene body methylation while low expressed genes tend to have lower levels (Fig. 10-2a). However, how gene body methylation in mammals contributes to gene expression is still unclear.

Remarkably, we found distinct differential methylation at gene bodies between up- and down-regulated genes. Whereas upregulated genes on average have lower levels of gene body methylation, downregulated genes tend to have higher levels of gene body methylation compared to the genomic average (Fig. 10-2b, Fig. S3). This result supports the hypothesis that higher levels of gene body methylation contributes to enhanced gene expression in mammals\textsuperscript{5}. For example, \textit{Cad} has low promoter methylation and high levels of gene body methylation but becomes downregulated in TKO cells (Fig. 10-2c), suggesting the downregulation may be due to loss of methylation in the gene body but not promoter. Our results indicated that loss of gene body methylation in these genes is associated with reduced expression. In contrast, upregulated genes show lower levels of gene body methylation and may be more likely regulated by promoter methylation, in accordance with the known role for DNA methylation. Indeed, cross-referencing methylation patterns with the recently available mouse methylome\textsuperscript{20} confirmed up-regulated genes have significantly greater promoter methylation and reduced gene body methylation compared to down-regulated genes (Fig. S4). Overall, our analysis uncovered a more direct association between gene body methylation and expression in mammals. In addition, it has been postulated that non-CG methylation in the gene body may also contribute to regulation gene expression (Lister et al. 2009). Although we found non-CG methylation patterns follow CG patterns (Fig. S3), we could not distinguish whether loss of CG or non-CG methylation resulted in reduced expression. Therefore, it is still unclear whether non-CG methylation is significant in regulating genes in mouse ESCs.
De-regulation of differentiation genes are associated with histone H3K27 trimethylation redistribution.

Although a previous study has shown that total histone abundance is similar between WT and TKO cells, it is still unclear whether histone localization is altered in the genome. To determine whether altered histone distribution in TKO ESCs contribute to gene deregulation, we first looked into genome-wide occupancy of H3K4me3 and H3K27me3, histone modifications involved in transcriptional activation and repression. We performed chromatin immunoprecipitation of H3K4me3 and H3K27me3 in TKO cells followed by high-throughput sequencing (ChIP-Seq). Using previously published chromatin maps, we compared genomic occupancy of H3K4me3 and H3K27me3 between WT and TKO ESCs.

Using a criterion of FDR < 0.1% for significant peaks, we identified 1kb windows that were significantly enriched for H3K4me3 or H3K27me3. Globally, nearly all genes enriched for H3K4me3 at the transcription start site (TSS) in WT cells remained H3K4me3 positive in TKO cells (Fig. 10-3a and 10-3b), suggesting H3K4me3 localization is mostly independent of DNA methylation. In addition, H3K4me3 enrichment outside the TSS was also mostly unchanged (Fig. 10-3b), including pericentromeric heterochromatin regions, which have been shown to be acquire H3K4me3 occupancy in hypomethylated mouse embryonic fibroblasts.

We therefore examined H3K27me3 re-distribution in TKO cells in relation to DNA methylation and gene expression. In sharp contrast to H3K4me3, we found prevalent genome-wide re-distribution of H3K27me3, most notably at the TSS (Fig. 10-3a and 10-3b), suggesting DNA methylation and H3K27me3 are closely coupled to each other. We found the majority of H3K27me3 positive genes often co-localized with H3K4me3, forming so-called bivalent
domains. Strikingly, 2071 of the 3563 (58%) bivalency genes in WT cells lose H3K27me3 occupancy in TKO cells (Fig. 10-3c), resulting in monovalent H3K4me3 domains in TKO cells. As expected, these 2071 genes were primarily associated with differentiation. Curiously, most genes that lose bivalency showed no significant gene expression change, though there was still a significant number of upregulated genes (n=241, p < 0.01). Gene ontology analysis on these set of genes reveal enrichment in transcription factors involved in various differentiation lineages including neurogenesis. Thus, a sizable portion of upregulated differentiation genes in TKO cells may be contributed by re-distribution of H3K27me3 at bivalent domains (Fig. 10-4a).

Those genes that are down-regulated in TKO cells tend to have only H3K4me3 positive domains in both WT and TKO cells, suggesting an indirect mechanism of gene repression in TKO cells. Consistently, methylome analysis revealed H3K4me3 positive genes also tend to have hypomethylated promoters as well as higher levels of gene body methylation (Fig. 10-4b). Because the chromatin environment is virtually unchanged for the set of H3K4me3 positive downregulated genes, loss of gene body methylation may be a more likely mechanism for reduced expression.

To determine relationships between DNA methyltransferases and the observed patterns in gene deregulation and H3K27me3 re-distribution, we stably re-expressed a single Dnmt (such as Dnmt1, Dnmt3a1, Dnmt3a2, and Dnmt3b1) in the TKO mESC background (Fig. S4). To determine whether de novo methyltransferases have a causal relationship with H3K27me3 occupancy, we performed H3K27me3 ChIP-Seq in TKO+Dnmt3a2 ESCs and used TKO+Dnmt1 as control. Interestingly, H3K27me3 occupancy was not restored in both reconstitution cell lines (Fig. 10-4c), suggesting that neither Dnmt1 nor Dnmt3a2 alone is sufficient for recruiting H3K27me3 to gene promoters.
Mass spectrometry of TKO reconstituted Dnmt cell lines showed DNA methylation levels are partially restored, indicating these enzymes are catalytically active. To determine the genomic localization of the methylcytosine marks, we performed RRBS in all TKO Dnmt reconstituted cell lines. In addition, we used combinations of single and double knockout mES cell lines to validate our results (Fig. 10-5a and Table S3). Globally, RRBS revealed that each methyltransferase has unique contributions to the total methylation level in mESCs (Fig. 10-5b). Single knockout or reconstitution of each de novo enzyme revealed that Dnmt3a and Dnmt3b share very similar methylation activity. Surprisingly, we were able to identify a small yet significant increase in CG methylation in TKO+Dnmt1 compared TKO mESCs, arguing that Dnmt1 has some de novo activity. Consistently, mESCs null of both Dnmt3a/3b also showed trace amounts of CG methylation greater than TKO cells (Fig. 10-5b). Dnmt1 de novo activity appear to be exclusively CG methylation as non-CG methylation did not show a significant different with TKO. However, both Dnmt3a and Dnmt3b showed comparable levels of non-CG methylation activity (Fig. 10-5b).

To examine methylation patterns in different genomic contexts, we plotted the average methylation level across genes and repeats. Consistent with global levels of methylation, we see proportional restoration of methylation at gene promoters and gene body (Fig. 10-5c). Significantly, we observe a well-defined methylation pattern across genes (i.e. select demethylation of proximal promoter and increasing gene body methylation), suggesting the Dnmt targeting mechanisms are unaltered. De novo methyltransferases appear to be predominantly targeted to gene body methylated since deficiency of Dnmt1 does not affect the gene body pattern, whereas deficiency in both Dnmt3a/3b results in loss of gene body methylation (Fig. 10-5c). We next examined methylation patterns at repeat elements and found
the majority of trends observed globally and at genic regions were consistent at various repeat elements. Strikingly, IAP elements showed high levels of methylation in TKO+Dnmt1 mESCs, suggesting Dnmt1 has de novo methylation activity targeted to IAP repeat elements (Figure 10-5d and Fig. S5). Importantly, the level of de novo methylation was comparable to single Dnmt3a or Dnmt3b TKO reconstitution, indicating Dnmt1 has comparable levels of de novo activity compared to the de novo enzymes. Taken together, our results revealed unique and shared genomic targets of each DNA methyltransferase.

**DISCUSSION**

In this study, we demonstrated that DNA methylation plays multiple roles in regulating global RNA expression. Using high-throughput RNA-Seq technology, we were able to identify a much larger number of deregulated genes than previous reports. This finding is consistent with other mRNA-Seq studies, which showed a much improved sensitivity in detecting gene de-regulation compared to microarray technology. Importantly, we learned that multiple RNA-Seq datasets from different laboratories show non-biological variations that potentially confounds the data analysis. This may not be a surprise since we and others have observed similar phenomena in microarray platforms 15,23. Because high-throughput sequencing technology is rapidly developing, new sources for variations will constantly need to addressed. For example, different methods for library construction and newer generation sequencing machines have inherently different biases than its predecessors. Technical variations may also come from varying RNA fragmentation efficiencies, resulting in differential sampling of transcriptomes, giving rise to discrepancies in identifying differentially expressed genes. Thus, correction of batch effects should increasingly become a routine part of quality control when analyzing multiple RNA-Seq experiments.
Nevertheless, our gene ontology analyses are largely consistent with previous studies concerning the up-regulation of differentiation-related genes, suggesting we are still able to measure biological differences between WT and TKO ESCs.

By combining DNA methylation profiling through global bisulfite sequencing with RNA-Seq, we confirmed that a subset of differentiation genes are direct targets for DNA methylation-mediated gene silencing. Interestingly, we also found a potential role for gene body methylation whereby down-regulated genes tend to have higher levels of gene body methylation than upregulated genes. We note that RRBS coverage has biases toward CpG rich regions, indicating the set of downregulated genes associated with high levels of gene body methylation are usually CG dense. Thus, heavy gene body methylation seems to contribute to gene expression for the set of genes with high CG density across the gene body.

Furthermore, our current study demonstrated that loss of DNA methylation somehow leads to re-distribution of H3K27me3, raising the possibility that DNA methylation plays a necessary and synergistic role in silencing differentiation genes with repressive histone modifications at both gene promoters and gene body. Most notably, we found loss of H3K27me3 at bivalency domains in TKO ESCs tend to de-repress differentiation genes. However, there is still a large portion of bivalency genes that resolve to monovalency but does not show change in gene expression. It is possible that these set of genes are poised for activation pending on endogenous differentiation cues that are still missing or they are actively inhibited by the intact pluripotency network in TKO cells. Some conflicting studies surround Dnmt interaction with the PRC2 complex which catalyzes, at least partly, H3K27 trimethylation. One study has shown that PRC2 associates with Dnmt1, Dnmt3a and Dnmt3b, though a separate study has shown that Dnmt3a can antagonize the PRC2 complex. Consistent with our study, another report has
shown that TKO mESCs have marked loss of H3K27me3 at the transcriptional start site. Here, we sought to determine causal relationships between Dnmts and H3K27me3 by re-expressing a single Dnmt in TKO mESCs. Our results indicated that single Dnmt1 or Dnmt3a expression alone is insufficient to restore H3K27me3. This would suggest that the Dnmts work in synergy to shape the epigenome.

Our unique TKO Dnmt reconstitution system identified unique and shared targets of each DNA methyltransferases. Most strikingly, we observed de novo methylation activity in Dnmt1 that can be traced to targeting the IAP repeat element. Two Dnmt1 structure papers have predicted that Dnmt1 due to the auto-inhibitory CXXC domain which proteins unmethylated DNA from spurious de novo methylation\(^{26,27}\). By contrast, hemi-methylated DNA antagonizes auto-inhibition and facilitates Dnmt1 maintenance activity. We and others have previously showed that DNA methylation also plays a contracting role in silencing the retrotransposable elements such as intra-cisternal A particles (IAPs) in ESCs and somatic cells. While DNA demethylation in differentiating somatic cells causes a huge induction of IAPs, it only triggers a minor change in the expression of endogenous retroviral elements in ESCs\(^{14,28}\).

In addition to mRNA expression and histone codes, other clues that TKO ESCs are in a state of up-regulation of differentiation genes are the downregulation of many lincRNAs, which have been shown to maintain pluripotency and suppress differentiation\(^ {17}\). A portion of lincRNAs have been shown to be direct targets of transcription factors such as Oct4, Sox2, and Nanog, which are concomitantly downregulated upon differentiation. However, Oct4, Sox2, and Nanog are highly expressed in TKO ESCs, suggesting the set of deregulated lincRNAs in TKO may act independent of the pluripotency network and instead are somehow indirect targets of DNA methylation.
Our current study suggests that the role of DNA methylation in gene regulation may be cell content-dependent. Potentially, the list of deregulated genes in TKO ESCs uncovered in this study may be specific for mouse ESCs and may be different from the deregulated genes in hypomethylated somatic cells. We suggest that the specific role for DNA methylation in gene regulation may depend on specific transcription networks in a particular cell type. Finally, alterations in DNA methylation or mutations in DNMTs have been associated with several human diseases including immunodeficiency, mental retardation, and cancer. To understand the disease mechanisms, it will be critical to understand the direct versus indirect effects of DNA methylation on the deregulation of transcriptome in these genetic disorders. Such information will be very useful for designing a therapeutic approach to correct the perturbed gene expression patterns in these disease cells.

**Material and Methods**

**ES cell cultures**

ESCs were maintained as previous described. RNA was isolated using Trizol (Invitrogen) while DNA was isolated using DNA lysis buffer, and then phenol:chloroform extracted.

**mRNA-Seq**

Total RNA were enriched for poly-adenylated tails using oligo-dT column (Qiagen). Sequencing libraries were constructed using the suggested protocol, followed by sequencing on Illumina Genome Analyzer IIx or HiScan sequencer according to manufacture protocol for single or paired-end of 76 cycles, respectively. Reads were mapped to genome using Burrows-Wheeler Aligner (BWA) with max of 3 mismatches. Only uniquely mapping reads were considered. We used DeSeq with parameters sharingMode=”fit-only” and fitType=”local”. Multiple
testing was corrected by implementing < 1% FDR criterion. Differential expression also required 1.5 fold cutoff and a minimum of 10 reads in either TKO or control library for mRNA analyses.

**ChIP-Seq**

Chromatin immunoprecipitation was performed as previously described and library construction was performed as indicated by manufacture protocols. Antibodies for H3K4me3 from and H3K27me3 were purchased from Millipore. To compare with previous published dataset, we down-sampled our TKO ChIP to similar number of reads. Peak calling was perform as previously described.

**Gene ontology analysis**

Functional enrichment was assessed using the DAVID database http://david.abcc.ncifcrf.gov/. The statistical threshold level for all gene ontology enrichment analyses was $P < 0.05$ (Benjamani and Hochberg corrected for multiple testing).

**Quantitative Reverse Transcription PCR**

RNA was DNase I treated (Invitrogen) and then quantified again. cDNA conversion was done using the iScript kit (Bio-Rad). Quantitative PCR was done on a MyIQ Thermocyler (Bio-Rad) using the Sybr Green Supermix (Bio-Rad).

**Bisulfite sequencing and DNA methylation analysis**

Reduced representation bisulfite sequencing libraries were constructed as previous described. Sequencing was carried out using the Illuma HiSeq2000 sequencer according to the manufacturer’s protocol. 100-nucleotide reads were generated and mapped to the mm9 genome using BS-Seeker. After methylation calling, post-filtering was performed as previously described. Briefly, reads with more than 3 unconverted cytosines in non-CGs contexts were considered as incompletely converted and discarded. Binomial tests were used to further
determine significance of a methylated cytosine, using the error rate (T to C mismatch) as probability of success. All downstream analyses were performed using scripts written in Matlab.

**Acknowledgements**

K.H. is supported by the California Institute of Regenerative Medicine Predoctoral Training Grant (TG2-01169). This work was supported by in parts by the National Institute of Health (NS072924, P01GM081621, and GM081621).
References


Karimi, M. M. et al. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell stem cell* **8**, 676-687


Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*


26 Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science 331, 1036-1040


28 Hutnick, L. K., Huang, X., Loo, T. C., Ma, Z. & Fan, G. Repression of retrotransposonal elements in mouse embryonic stem cells is primarily mediated by a DNA methylation-independent mechanism. J Biol Chem 285, 21082-21091,


FIGURE LEGENDS

Figure 10-1. Global mRNA deregulation in TKO cells. a) Scatterplot of average read counts for each gene in WT and TKO ESCs. b) bar graph showing overlap of our identified gene list with other previous reports as well as validation using qRT-PCR. qRT-PCR was done using multiple WT and TKO ESC lines and average and standard deviation was used to determine statistical significance. Note that the TKO cells described by Fouse et al. (2008) are Dnmt3a−/−; Dnmt3b−/−; Dnmt1 knockdown. c) gene ontology analysis of up and down regulated genes using software from DAVID. d) genome browser view of all data, we chose 1 representative WT and TKO pair line for simplicity.

Figure 10-2. Global methylation patterns of deregulated genes in wild-type ESCs. a) methylation metaplot analysis of all genes ranked by RPKM and separated into quantiles. X-axis represents distance relative to size of gene body (0 = transcription start, 100 = transcription end). b) promoter or gene body methylation across all genes, significant was tested using student t-test. We defined promoter as 1kb region flanking the TSS, and gene body as CpG rich-regions between TSS and transcription termination site (TTS). c) genome browser view demonstrating downregulated gene with low promoter methylation and high gene

Figure 10-3. H3K4me3 and H3K27me3 distribution patterns in WT and TKO ESCs genome. a) metagene plot describing number of significant windows with either H3K4me3 or H3K27me3 enrichment. b) heatmap showing histone enrichment around TSS. c) bar graph showing specific histone modification transition between WT and TKO TSS. D) chromosome wide view of chromosome X showing global H3K4me3/H3K27me3 patterns in WT and TKO ESCs.
Figure 10-4. Methylome and gene expression analysis of histone re-distribution bivalent domains. A) Pie graph showing distribution of histone occupancies in WT for the set of up- and down-regulated genes. b) methylome analysis of average methylation levels across genes with different histone occupancies in WT cells. c) Bar plot showing number of gene promoters positive for H3K27me3.

Figure 10-5. Unique and shared targets of individual DNA methyltransferases.

a) schematic of color representation for presence of absence of individual Dnmts. b) bar plot showing methylation levels in different cytosine contexts. c,d) methylation metaplot analysis of all genes or IAP repeat element.
Figure 10-2.
Figure 10-3.
Figure 10-4.
Figure 10-5.
Chapter 11
Concluding Remarks

In this dissertation, I worked in collaborative efforts to identify key insights and interpretations from large datasets involving a variety of cell types. These sets of studies have highlighted the wide utility of high-throughput technologies in different biological contexts.

In one set of analyses, I identified distinguishing features of a single cell-type relative to other cell types. In two separate studies, I aimed to determine whether human ES and iPS cells were different or not. Through DNA methylation profiling and weighted gene co-expression network analysis, we concluded that ESCs and iPSCs are not entirely identical, and identified a robust CG methylation signature as well as co-expression modules that distinguish them. However, in both studies it is still unclear the functional consequence of these distinguishing marks, and will require experiments on the bench to investigate further. Additionally, I identified another set of signature genes in retinal pigmented epithelium and corneal endothelial cells using global gene expression data. These gene expression profiles may be useful biomarkers for determining the identity of a cell, such as during stem cell differentiation. Together, this work provides useful markers to evaluate the maturity and quality of pluripotent stem cells and their differentiated products.

Along the line of monitoring the success of stem cell derived tissue, I have worked together with a team of scientists who have sampled different stages of stem cell differentiation and development with the goal to improve the resolution and understanding for the molecular changes during the process. Using the RPE differentiation paradigm, we profiled both miRNA and DNA methylation patterns at four separate time points and revealed dynamic regulation of both epigenetic mechanisms. The information from the intermediate stages further helped refine future work in determining the progress of stem cell differentiation. In a separate study, I
analyzed the transcriptional landscape of early human embryos ranging between mature oocytes and morula embryos via single-cell RNA-sequencing. This technique is exceptionally powerful because we were able to maximize the transcriptome information—in terms of sample size, genomic coverage, and base-resolution—from these precious embryos. Importantly, this study demonstrates the feasibility to assay global molecular patterns in a rare population of cells, opening new opportunities for expanding genomic and epigenomic studies in the future.

The final chapters of this dissertation aimed to understand the role of DNA methyltransferases in shaping the methylome of pluripotent stem cells. We utilized two separate model systems; first in the case of naturally occurring mutations in the DNMT3B enzyme that lead to ICF Syndrome and second in the transgenic mouse models null of all three DNA methyltransferases. These studies have revealed that each DNA methyltransferases has unique contributions in targeting the genome. Significantly, our results suggest that although human and mouse DNMTs share many similarities, they also diverge in genomic targets and enzyme redundancy. For example, whereas DNMT3B seem to be crucial for non-CG methylation, loss of Dnmt3b in mouse appears to be mostly compensated by Dnmt3a and vice versa. Thus, mouse Dnmts seem to have more functional redundancy compared to their human counterparts. Overall, these studies provide a comprehensive and high-resolution insight into the consequence of de-regulated Dnmt in mammalian pluripotent stem cells.

Taken together, this dissertation examines many different model systems in the context of high-throughput technologies. The advantages of using high-throughput technology compared to previous single-gene methods is the ability to have comprehensive view of the molecular features in the cell of interest in a cost-effective and less time-consuming way. In addition, sampling a large amount of genes in a single experiment facilitate stronger statistical conclusions
with high confidence. For example, it is now commonly accepted that statistics are reported in terms of a post-hoc false-discovery rate (FDR). These methods compensate for multiple testing (i.e. the chance that a significant p-values is seen increases as sample size increases) and provide an intuitive understanding of how often the number of positive hits are expected to be false. At the same time, datasets have become more complicated and new methods are constantly being developed and optimized to simplify large datasets to draw meaningful conclusions. For example, Chapter 3, 4, and 6 employed weighted gene co-expression networks (WGCNA), which considers that all genes may have some correlated relationship with another set of genes, so-called “guilt-by-association”. The advantage of using WGCNA over traditional methods is that it alleviates the multiple testing problem as mentioned above and generally leads to more reproducible findings\(^1\). Gene networks are thought to be more stable because the overall function of the network can persist while individual gene expression within the network may fluctuate\(^2\). Together, high-throughput technologies have opened comprehensive snapshots of cellular features and enabled increased statistical power.

The work in this dissertation spans a period of four years, yet it has seen major shifts in the preferred choice of high-throughput methodology. In particular, high-throughput sequencing (HTS) appears to be becoming the preferred platform compared to microarrays. This is because sequencing has advantages in unbiased and unrestricted genome assessment at high resolution. The advantages of HTS may outweigh the microarray platform as the former become more cost-effective than the latter. Indeed, many of the major bodies of institutions have formed consortiums such as ENCODE and the Epigenomics Roadmap Project that are dedicated to creating a comprehensive resource by high-throughput sequencing all useful biological models. This will undoubtedly be an invaluable resource that will facilitate many future studies.
Looking forward, as these consortia and others exhaustively and massively sequence all the different biological systems, single-cell sequencing may be the next frontier of genomic and epigenomic studies\textsuperscript{3}. Chapter 8 highlights the power of single-cell analysis using a comprehensive RNA-sequencing approach, though previous studies in single-cell analysis using the Fluidigm system have drawn meaningful conclusions as well, notably that cell populations appear to be more heterogeneous than previously thought\textsuperscript{4-6}. Thus, some of the current interpretations in the stem cell field may also be confounded by cell population heterogeneity, and future single-cell analysis may resolve some of the controversial claims.

The field of epigenomics has been made possible by high-throughput technologies, and this has allowed us to procure genomic maps of the epigenetic landscape. In the process, we have opened a Pandora’s Box of complexities that will probably take years to fully decipher. The deluge of data will ultimately rely on adequate manpower to analyze all the data in an integrative and intuitive manner to uncover meaningful biology. Importantly, key collaborations between bioinformaticians and biologists will garner improved efficiency on both sides. Remarkably, current estimates indicate that sequencing costs are actually outpacing Moore’s law since the completion of the human genome project\textsuperscript{7}. This will certainly make studying epigenomics orders of magnitude more affordable and facile in the future. Aided by increasingly optimized genomic methods, many more waves of epigenetic breakthroughs will surely be discovered and fields such as stem cell biology will certainly benefit from the wealth of epigenetic information.
References


