THE UNIVERSITY OF CHICAGO

CHARACTERIZING INTER-INDIVIDUAL REGULATORY VARIATION USING INDUCED PLURIPOTENT STEM CELLS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

NICHOLAS ELI BANOVICH

CHICAGO, ILLINOIS

AUGUST 2016

Table of Contents

LIST OF FIGURES ..... **III**

LIST OF TABLES ..... **IV**

LIST OF SUPPLEMENTAL FIGURES ..... **V**

LIST OF SUPPLEMENTAL TABLES ..... **VI**

ACKNOWLEDGMENTS ..... **VII**

ABSTRACT ..... **X**

CHAPTER 1: INTRODUCTION **1**

CHAPTER 2: METHYLATION QTLS ARE ASSOCIATED WITH COORDINATED CHANGES IN TRANSCRIPTION FACTOR BINDING, HISTONE MODIFICATIONS, AND GENE EXPRESSION LEVELS **7**

2.1 ABSTRACT 8
2.2 INTRODUCTION 9
2.3 RESULTS 12
2.4 DISCUSSION 25
2.5 MATERIALS AND METHODS 29
2.6 APPENDIX A: SUPPLEMENTARY MATERIALS 38

CHAPTER 3: GENETIC VARIATION, NOT CELL TYPE OF ORIGIN, UNDERLIES THE MAJORITY OF REGULATORY DIFFERENCES IN IPSCS **44**

3.1 ABSTRACT 45
3.2 INTRODUCTION 45
3.3 RESULTS 48
3.4 DISCUSSION 58
3.5 MATERIALS AND METHODS 60
3.6 APPENDIX B: SUPPLEMENTARY MATERIALS 71

CHAPTER 4: HUMAN INDUCED PLURIPOTENT STEM CELLS: A POWERFUL MODEL TO INVESTIGATE INTER-INDIVIDUAL REGULATORY VARIATION ACROSS CELL TYPES **83**

4.1 ABSTRACT 84
4.2 INTRODUCTION 85
4.3 RESULTS 87
4.4 DISCUSSION 102
4.5 MATERIALS AND METHODS 104

CHAPTER 5: DISCUSSION **116**

REFERENCES **123**

LIST OF FIGURES

Figure 2.1 meQTLs identified in LCLs 14
Figure 2.2 meQTLs are enriched for QTLs from other regulatory phenotypes ... 17
Figure 2.3 A single SNP is associated with coordinated change in multiple regulatory phenotypes 19
Figure 2.4 SNPs disrupting TF binding sites drive changes in DNA methylation 21
Figure 3.1 Study Design. 49
Figure 3.2 Hierarchical clustering and principal components analysis. 51
Figure 3.3 Differential Methylation and Gene Expression Between the Four Cell Types (L-iPSC, F-iPSC, LCLs and fibroblasts). 54
Figure 3.4 Contribution of Individual Differences Versus Cell Type of Origin to Methylation and Expression Levels. 57
Figure 4.1 Quality control of iPSC lines 88
Figure 4.2 Regulatory variation is lower in iPSCs 90
Figure 4.3 Properties of eQTLs across cell types 93
Figure 4.4 Regulatory annotations driving cell type specific and shared eQTLs 96
Figure 4.5 cell type specific caQTLs 97
Figure 4.6 An iPSC specific caQTLs that drives cell type specific changes in expression 98
Figure 4.7 iPSC-CMs replicate expression variation from primary heart tissue 100
Figure 4.8 iPSC-CMs enable the study of heart disease phenotypes 101

LIST OF TABLES

Table 2.1 Associations between QTLs for other regulatory phenotypes and DNA methylation 18
Table 2.2 Associations between SNPs disrupting TF binding sites and DNA methylation within 500bp of the binding site 23
Table 2.3 DAVID analysis of meQTLs implicated in GWAS 24
Table 4.1 Gene ontology enrichment of heart specific processes in iPSC-CM eGenes 101

LIST OF SUPPLEMENTAL FIGURES

Figure S2.1 Scatterplot of CpG methylation levels estimated from the Illumina array and from whole genome bisulfite sequencing. 38
Figure S2.2 A boxplot of distances from methylation probe to transcription start site for eQTL/meQTLs. 39
Figure S2.3. QQ-plot of associations between SNPs implicated in GWAS studies and DNA methylation. 39
Figure S2.4. Distributions of methylation levels 40
Figure S2.5. Affect of normalization on meQTL calls 40
Figure S2.6. Enrichment of QTLs under different normalization procedures 41
Figure S2.7. PCA plots 42
Figure S3.1 Quality control of iPSCs. 71
Figure S3.2 Quality control of iPSC Lines. 72
Figure S3.3 Quality control of iPSCs. 73
Figure S3.4 Quality control of iPSCs. 74
Figure S3.5 DNA methylation density plots. 75
Figure S3.6 Array data normalization. 76
Figure S3.7 Probe inclusion scheme. 77
Figure S3.8 Hierarchical clustering. 78
Figure S3.9 Principal components analysis (PCA). 79
Figure S3.10 Heatmap of DM loci. 80
Figure S3.11 DE tests in gene subsets. 81
Figure S3.12 Differential methylation with single L-iPSC replicate. 82

LIST OF SUPPLEMENTAL TABLES

Table S2.1 Enrichment analysis of probes location 42

ACKNOWLEDGMENTS

I want to begin by thanking my advisor Yoav Gilad. He is an amazing scientist and I am extremely grateful that I got to spend the last four and a half years working with him. His mentorship style is very “direct” which took some getting used to, but I have benefited greatly from his candid assessment of my ideas and work. Even though Yoav is an extremely busy individual (conference travel, grant writing, parenting, etc.) he was always available to talk. I frequently took advantage of his open door policy and had brief meetings with him multiple times a week. In addition to my professional relationship with Yoav I enjoyed a great personal relationship with him as well. We quickly bonded over being parents and many of our meetings took long detours to talk about our kids. I will be forever grateful for the time I spent in the Gilad lab! I also want to say a special thank you to someone I considered a co-mentor, Jonathan Pritchard. I was able to spend a year and a half working closely with Jonathan before he moved to Stanford. Because of this time I was able to maintain a close relationship with Jonathan that included multiple visits to Stanford as well as many video calls! Working with Jonathan has been a great joy. There is nothing quite like sitting in a two hour meeting with Jonathan really dissecting a question and watching his mind work. I have rarely met a senior scientist so enthusiastic about the research he is doing.

I was fortunate to work with an amazing group of people in the Gilad lab. I want to give a special thanks to Emily Davenport, John Blischak, Darren Cusanovich, Irene Gallego Romero, Jack Degner, Roger Pique-Regi, Bryce Van

De Geijn, Graham McVicker and Sidney Wang. These lab members not only provided an immense amount of scientific and professional support, but also made the lab feel like home when I first arrived. This group of people was amazingly welcoming and took me under their wing. I also want to give a special thanks to Michelle Ward and Po Yuan Tung who have been amazing colleges and friends over the past two-three years. I don't know what I would have done without their support and collaboration. They have truly made my final two years an amazing experience. Finally I want to thank Yang Li, Anil Raj, Courtney Burrows, Jonathan Burnett, Athma Pai, Amy Mitrano, Brett Engelmann, Bryan Pavlovic, Sammy Thomas, Marsha Myrthil, Julien Roux, Sebastian Pott, and Claudia Garcia who all contributed directly to my projects either directly or indirectly.

I want to thank my family, especially my wife Katie. She moved to Chicago with me without question even though we had no family or friends here. Since moving here we have had two kids, Emery and Greyson, and she has been an amazing. Without her my life would be in disarray. I can't imagine anyone else but you. I also want to thank my parents, Erin and Tony, who have always believed in me, even when times were tough. I am honored to be their son and nothing makes me happier than seeing the pride in their eyes. Thank you guys for always loving me. I also want to thank my little sister Zoë. She is a warrior and an amazing person, I couldn't be prouder of her. I want to thank Eric Owen, Terry Murphy, and Bryan Wilcox for being a constant support in my life. Without you guys and Bill I would have nothing. I also want to thank Anthony Hilario,

Mackay Pierce, and David Williams (RIP). Finally, I want to thank my Grandfather who passed away on July 8th 2016 during the preparation of this document. I will be his first grandchild to receive a PhD.

ABSTRACT

As researchers have sought to understand the genetic architecture of complex traits, including disease, it has become apparent that the majority of the signal originates outside protein coding regions of the genome. These results, obtained through many genome wide association studies (GWAS), have led most to conclude that changes in the regulation rather than structure of genes is the driving force behind variation in complex traits. Therefore, many hundreds if not thousands of genetic variants with small effects drive variation in complex traits, making it difficult to identify meaning genetic variants. This has led many researches to focus on the effects of genetic variation on gene regulation as an intermediate phenotype. Here I present three works focused on improving our ability to understand the mechanisms underlying inter-individual variation in gene regulation and building a better systems to study these phenomena in disease relevant cell types. In my second chapter I will describe the effect of genetic variation on changes in DNA methylation levels and how these changes result in coordinate changes in histone modifications, transcription factor binding, and gene expression. My third chapter will focus on testing the fidelity of a new system, induced pluripotent stem cells, in which we can study gene regulation. Finally my fourth chapter will focus on characterizing inter-individual variation in gene regulation across three cell types (induced pluripotent stem cells, cardiomyocytes derived from induced pluripotent stem cells, and lymphoblastoid cell lines from which the induce pluripotent stem cells were derived).

CHAPTER 1: INTRODUCTION

Understanding how heritable genetic variation contributes to inter-individual variation in phenotype is one of the major goals of Human Genetics. In particular, elucidating the genetic mechanisms underlying phenotypic traits has great potential to better predict and treat disease Phenotypic traits in are generally broken down into two major classes, simple or complex. Simple traits are those where a single locus contribute to the trait, for example, sickle cell anemia or Tay-Sachs disease. These traits are characterized by a binary phenotype (given complete penetrance) and generally the alleles driving the traits disrupt the protein coding sequence of a gene or large segments of a chromosome. Complex traits are shaped by many loci. These traits are usually thought of a quantitative – i.e. rather than the trait being present or absent a distribution of the trait exists in the population. One clear example is that of height, which demonstrated to be quantitative as early as 1914 [1]. Initial efforts to map both complex and simple traits relied upon linkage studies [2]. This method, first described by Botstein et al in 1980 [2] relies on family pedigrees where the trait and alleles are liked as they segregate in a family. The use of linkage studies was quite successful identifying the loci underlying simple traits – due to a single segregating locus with large effects and frequently high penetrance [3-8]. However, mapping loci underlying complex traits proved much less successful [9,10], in large part due to the high number of variants underlying complex traits and the low resolution of linkage mapping. Fortunately, in the past 15 years there have been rapid advances in sequencing technologies allowing new methods to be developed to map complex traits.

The mapping of a full human genome [11] along with the advent of high throughput sequencing and array based genotyping allowed researchers to begin performing unbiased genome-wide scans for genetic variations associated with complex traits. Rather than searching for genetic variants that segregate within a family, these studies typically employed a case control design (of unrelated individuals) and test for differences in allele frequencies of all assayed genetic variants between groups.

Genome wide associations studies or GWAS promised to be more successful than linkage analysis and many hoped they would uncover the majority of the variation underlying complex traits [3]. Unfortunately, the complexity of these traits was greater than realized. For example, a recent study performed a GWAS for body mass index (BMI) in over 300,000 individuals. Their analysis identified 97 loci associated with BMI at genome-wide significance ($P < 5 \times 10^{-8}$) [12], which account for only around 3% of the BMI variation. The inability of GWAS to explain a high proportion of the expected heritability of complex traits in not unique to the study of BMI; indeed, these finding have been replicated across many complex traits [3,13-16]. The results from these studies and others suggest there could be thousands of loci contributing to variation in complex traits. Unfortunately, this genetic architecture makes it exceedingly difficult to elucidate the genetic underpinnings of complex traits using simple associations alone. This is not to say that GWAS have been a total failure. Indeed, GWAS have provided valuable insights into the general principles of genetic architecture underlying complex traits. One of the major findings from

GWAS is that the vast majority of genetic variants associated with complex traits do not reside within the protein coding regions of the genome, but rather in regions responsible for the regulation of gene expression [3,13,17]

Since genetic variants implicated in GWAS are often intergenic it is difficult to dissect the mechanism by which the variant acts on the trait of interest. Many studies suggest that the genetic variant acts on the nearest gene or genes [3], but these assumptions are often incorrect and can lead to researchers to follow up on incorrect genes [18]. Thus, a number of groups have set out to gain a better understanding of how genetic variation within the non-coding regions effects gene expression. Beginning in 2005 it was shown that inter-individual variation in mRNA levels could be mapped to genetic variants by performing association studies between genotype and gene expression levels [19]. These studies benefit from a reduced multiple testing burden by only considering genetic variants putatively acting in cis (within 100kb of a gene). Since 2005 there has been tremendous progress mapping genetic variation that is associated with changes in gene expression and elucidating the mechanisms by which genetic variants affect gene expression.

It has new been well demonstrated that genetic variation is a major driver of inter-individual variation in gene expression [20-33]. These results suggest that nearly every gene has at least one genetic variant affecting expression levels (eQTLs). In an effort to further dissect the mechanisms driving changes in gene expression many have begun instigating the effect of genetic variation on other aspects of gene regulation. Specifically, there have been association studies

performed between genetic variation and DNA methylation levels [27,34-36], chromatin accessibility [37] and histone modifications [38-41]. These studies demonstrate the ability of a genetic variant to affect many regulatory phenotypes in concert. Importantly, this work has found that disrupting transcription factor binding sites (TFBS) has the ability to alter chromatin function and result in expression changes [36,40].

It has been demonstrated that jointly modeling genetic variation associated with expression and complex traits leads to an increase in power to identify variation associated with a complex trait and a better understanding of the underlying biology [42]. However, one major weakness detracting from the majority of eQTLs studies is the inability to perform such association studies in disease relevant tissues. Much of the work already mentioned was performed in immortalized cell lines. There are two major reasons for this shortcoming. First, cell lines are easy to maintain. They can be frozen and thawed indefinitely and used repeatedly for numerous studies. Additionally, it is feasible to obtain the millions of cells necessary for many of these analyses. Second, it is difficult to obtain primary tissue from living individuals, both practically and ethically. Often the tissues we are most interested in are critical for life and therefore cannot be sampled. A major effort by the GTEx consortium has collected post-mortem tissue from thousands of individuals and identified eQTLs across these tissues [32]. This represents a major advance to the field but has it's own set of caveats. Namely, this tissue is finite and it is static. This results in a limited number of regulatory phenotypes that can be assayed (at this time only mRNA levels have

been characterized). Additionally, no perturbations can be performed on the tissues; thus only a single snapshot of steady state gene expression levels at the time of death is obtainable.

A promising technology has emerged, which may aid in overcoming the shortcomings of current models. Namely, the discovery that human somatic cells can be reprogrammed into a pluripotent state [43-45] and then be differentiated [46] into multiple somatic lineages, has the potential to provide access to a wide range of cell types from practically any donor individual. Since the initial discovery of induced pluripotent stem cells (iPSCs) they have been used in a wide range of studies, mainly to model disease in vitro or rescue disease phenotypes in vivo [47-60]. However, their usefulness as a model system to study human phenotypes remains debated [61-63].

The major goal of this thesis is to demonstrate the usefulness of the iPSC model in the study of human traits, specifically inter-individual variation in gene regulation. I began by studying the effect of genetic variation on DNA methylation levels (meQTLs) in immortalized lymphoblastoid cell lines (LCLs), which will be transformed into iPSCs in the 4th chapter. Next, I set out to identify the major sources of gene expression and DNA methylation variation in iPSCs. Finally, I generate a large panel of iPSCs from a West African population, the Yoruba. I characterize genetic variation associated with gene expression in iPSCs and demonstrate the usefulness of iPSCs and iPSC-derived cell types as a model to study human traits.

CHAPTER 2: METHYLATION QTLS ARE ASSOCIATED WITH COORDINATED CHANGES IN TRANSCRIPTION FACTOR BINDING, HISTONE MODIFICATIONS, AND GENE EXPRESSION LEVELS

2.1 Abstract[1]

DNA methylation is an important epigenetic regulator of gene expression. Recent studies have revealed widespread associations between genetic variation and methylation levels. However, the mechanistic links between genetic variation and methylation remain unclear. To begin addressing this gap, we collected methylation data at ~300,000 loci in lymphoblastoid cell lines (LCLs) from 64 HapMap Yoruba individuals, and genome-wide bisulfite sequence data in ten of these individuals. We identified (at an FDR of 10%) 13,915 *cis* methylation QTLs (meQTLs)—i.e., CpG sites in which changes in DNA methylation are associated with genetic variation at proximal loci. We found that meQTLs are frequently associated with changes in methylation at multiple CpGs across regions of up to 3 kb. Interestingly, meQTLs are also frequently associated with variation in other properties of gene regulation, including histone modifications, DNase I accessibility, chromatin accessibility, and expression levels of nearby genes. These observations suggest that genetic variants may lead to coordinated molecular changes in all of these regulatory phenotypes. One plausible driver of coordinated changes in different regulatory mechanisms is variation in transcription factor (TF) binding. Indeed, we found that SNPs that change

[1] Citation for chapter: Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, et al. (2014) Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLoS Genet 10(9): e1004663. doi:10.1371/journal.pgen.1004663

predicted TF binding affinities are significantly enriched for associations with DNA methylation at nearby CpGs.

2.2 Introduction

Changes in gene expression levels are important contributors to phenotypic variation in human populations [20-30]. One way in which gene expression levels may be altered is through changes in chromatin function [35,38-40,64-67]. Recent studies have focused on identifying genetic variants that impact chromatin function [65,68] by studying inter-individual variation in DNase I sensitivity, a general indicator of chromatin accessibility [37], as well as a variety of histone modifications [38-41]. A single genetic variant was often found to be associated with coordinated changes in multiple molecular phenotypes, including chromatin accessibility, nucleosome positioning, chromatin modifications and gene expression levels [38-40]. In many cases of coordinated changes, the associated genetic variants seem to act through the disruption of transcription factor binding sites [38-40]. This body of work highlights the value of using multiple molecular phenotypes to understand the connection between genetic variation and gene expression. One important epigenetic mark not considered by these recent integrated studies is DNA methylation.

DNA methylation refers to the addition of a methyl group to cytosine nucleotides. In vertebrates, DNA methylation primarily affects cytosines that are immediately 5' to guanines, i.e., CpGs. Appropriate methylation is essential for

development and cellular differentiation [69-71]. Changes in DNA methylation levels have been linked to a number of diseases including tumorigenesis, [72,73] age-related defects [74,75] and mental disorders [76,77]. Typical array-based methylation assays provide a single measurement for each CpG site, which is interpreted to reflect the proportion of cells in which a given site is methylated. In general, this measurement was found to have a bimodal distribution across sites [35,78-80], which is believed to indicate that most sites are either methylated or unmethylated in nearly all cells in a given tissue or culture. Some measurements, however, are intermediate [79] (we refer to these as 'intermediate methylation levels'), which could either reflect methylation in a subset of cells or just in a single allele (one chromosome) in each cell. Most unmethylated CpGs are within CpG islands (CGIs), namely regions in the genome in which many CpGs are located in close proximity [79,81,82]. CGIs account for a small proportion of CpGs in the genome but they tend to be located near transcription start sites (TSSs). The methylation levels of CGIs are generally negatively correlated with the expression levels of nearby genes [35,79,81-83], an observation that led to a common early belief that DNA methylation was primarily a repressive epigenetic mark.

A number of studies have shown that genetic variation is often associated with quantitative changes in methylation levels [27,34,35,84,85]. Early QTL studies focused on methylation data from relatively few CpGs with a heavy bias towards promoter regions. A more recent study that used a comprehensive array platform considered genome-wide patterns and reported over 20,000 methylation

QTLs (meQTLs [34]). A number of meQTLs were also shown to be associated with changes in gene expression level (namely, these meQTLs are also classified as eQTLs) [27,34,35], although it is not clear whether the methylation changes are a cause or consequence of the gene expression changes [34]. Interestingly, in contrast to the early belief that methylation is primarily associated with repression, both direct and inverse correlations between methylation and gene expression levels have been observed. This suggests that the relationship between DNA methylation and gene expression levels may depend on the genomic context of the CpG [27,34,35].

In general, the mechanisms by which DNA methylation levels are being regulated remain unclear. One likely pathway is through coordination between DNA methylation and chromatin modifiers. For example, H3K4 methyltransferase is recruited by CFP1, which binds to unmethylated CpG islands [86]. In turn, H3K27me3 and DNA methylation have been shown to have mutually exclusive gene silencing functions, in at least some cases [87,88]. There is also limited evidence that TF binding may be associated with nearby changes in DNA methylation. For example, the insertion of a CTCF binding site was shown to cause changes in methylation levels near the insertion site (presumably due to the binding of CTCF) [80,89]. Less direct evidence comes from observations that TF binding sites are enriched in differentially methylated regions (DMRs) between individuals and cell types [90]. However, it is still unclear how frequently changes in TF binding affect the DNA methylation levels of nearby CpGs. It is also unclear whether this is a property that is associated with the binding of most

TFs or only a selected few. More generally, there has not yet been a broad examination of coordination between meQTLs and other molecular phenotypes.

In the current study, we therefore examined associations and correlations between genetic variation, DNA methylation, and multiple additional cellular regulatory phenotypes. We focused on a panel of Yoruba HapMap lymphoblastoid cell lines (LCLs), which have been extensively characterized in previous work. In addition to the methylation data we collected for the present study, genomic sequences are available for the majority of these lines [37], as well as RNA sequencing data and DNase I sensitivity profiles [37]. Histone modification data (profiles for H3K4me1, H3K4me3, H3K27ac, H3K27me3P) and PolII ChIP-seq data are also available for a subset of these lines [40].

2.3 Results

We measured methylation levels in 64 Yoruba LCLs using the Illumina Infinium HumanMethylation450 array, which assays methylation levels at roughly 450,000 cytosines, the majority of which are in CpGs. Probes on this array particularly target CpGs near transcription start sites, including CpG islands and CpG shores. As a first step in our data processing, we excluded array probes that did not uniquely map to the human genome as well as probes that overlapped a known sequence variant (see Methods). After these filtering steps we retained methylation measurements from 329,469 probes. As was suggested in previous studies [34,91,92], we quantile-normalized the data to a standard normal within each individual and across probes (though we considered the

effects of alternative normalization approaches; see Methods). To account for unobserved confounders we performed principal component analysis. We found that removing four principal components maximized our power to identify meQTLs. Further details on the data processing, normalization, and tests for the effect of confounders are provided in the Methods. In addition to the array data from 64 individuals, we also collected low-coverage whole-genome bisulfite sequencing data from a subset of ten individuals (median genomic coverage 2.4x; see Methods).

Mapping methylation QTLs

We first examined the association between genetic variation and differences in methylation levels across individuals. For this analysis, we considered only the array data (because we performed whole-genome bisulfite sequencing in only ten individuals). We used previously collected and imputed [37] genotype data for the 64 individuals from the HapMap and 1000 Genomes Projects [93,94]. We focused on proximal (putatively *cis*) associations between genotypes and DNA methylation levels by considering, in each case, genetic variation within a 6 kb region centered on the genomic location of a methylation probe on the array. This window size was chosen because smaller and larger windows yielded fewer significant associations at a given FDR. At an FDR of 10% we identified 13,915 CpG sites with at least one *cis* meQTL (Fig. 1A).

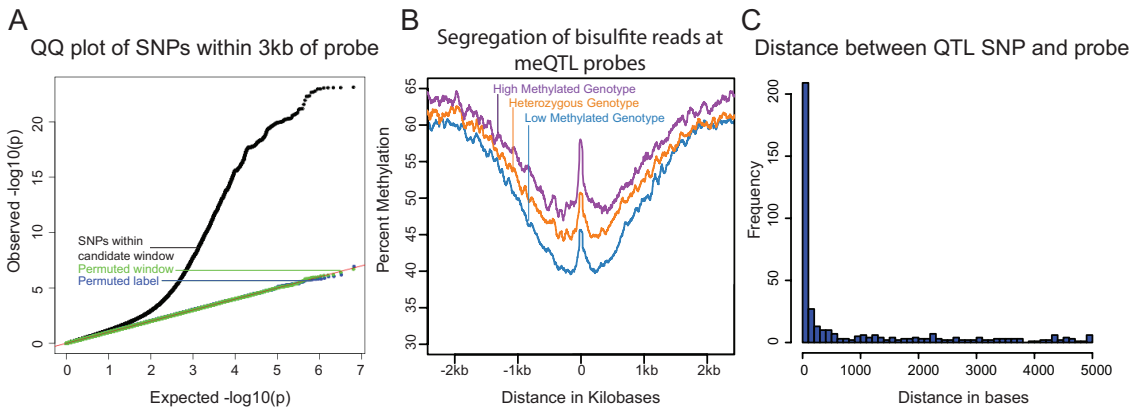

Figure 2.1 meQTLs identified in LCLs

A) QQ plot of –log10 p-values for testing the null of no association between methylation levels measured by all probes that passed our quality filters, and all SNPs within 3 kb of these probes. Data for SNPs within the candidate window are in black; negative control SNPs for which we chose a random 6 kb window elsewhere in the genome are in green; SNPs with the genotype labels permuted are in blue. B) Average methylation levels estimated using the bisulfite sequence data at meQTL probes, segregated by meQTL genotype. C) Histogram showing the distribution of distances between meQTL SNPs and the associated methylated sites in base pairs, for meQTLs where there is a single most likely causal site.

When multiple SNPs were significantly associated with methylation levels at a given site, we only considered (for the purpose of counting the overall number of meQTLs) the single most significant association. Since the methylation data measured by nearby pairs of probes are frequently correlated, we wondered whether this analysis might overstate the number of independent meQTL signals. To address this, we examined pairwise correlations of data from all probes located within 5 kb of each other. We found that data from only 203 or 520 of the associated probes (normalized or untransformed data, respectively) are significantly correlated (Pearson Correlation and a T-test; $P < 0.05$) suggesting that the reported number of independent meQTL is not substantially inflated by correlation of the methylation data across nearby probes.

We next used the genome-wide bisulfite sequencing data to provide a general validation of meQTL associations that were identified using the array data (Fig. 1B), as well as to investigate whether meQTLs are generally associated with changes in methylation at a single CpG or a larger region. In general, we observed a high correlation between the estimates of methylation levels based on the array data and the estimates of methylation levels based on the whole genome bisulfite sequencing ($R = 0.93$; Fig. S1). We note that the read depth and sample size of the bisulfite sequencing data set are insufficient to allow for validation of individual meQTL. Instead, we aggregated the sequence data by considering the centers of probe locations whose methylation data are associated with meQTLs (see Methods for more details). Using that approach, we found a clear difference in methylation level across meQTL genotypes. In addition, we observed a broad-scale association of meQTL genotypes with methylation levels over a region extending between 1.5 and 2 kb in either direction from the methylation loci originally probed by the array. This result indicates that multiple CpGs within a local region are often associated with a single meQTL.

We sought to estimate the typical distance between meQTLs and the location of associated methylated sites (based on the genomic location of the array probes). This analysis is complicated by the fact that, due to LD, it is often unclear which site is causal for any given meQTL. We thus focused on a subset of associations that are more likely to be causal, namely on 409 meQTLs that are the only strongly associated loci within 5 kb of the methylated site (see Methods).

Our approach does not provide direct evidence that these are indeed causal sites, but without additional experimental data (namely, using only the meQTL mapping framework), it is likely the best approach to obtain a subset of loci that is enriched with true causal associations [26,30]. These 409 meQTLs are generally located very near the associated methylation site (the median distance is 76bp; Fig. 1C), with only 52 (13%) of the putatively causal meQTLs located more than 3 kb away from the methylated site.

We then explored the distribution of methylated sites that are associated with meQTLs in the context of other *cis*-regulatory annotations. Using the chromatin state annotations from Ernst et al. [95], we classified the genomic regions containing the assayed methylated sites as insulators, enhancers, or promoters (see Methods). Compared to the distribution of all assayed methylation sites, we found a relative depletion of sites associated with meQTLs at promoters (chi-square test; $P < 10^{-15}$), and an enrichment of such sites at insulators (chi-square test; $P < 10^{-5}$) and enhancers (chi-square test; $P < 10^{-9}$; Table S1), consistent with previous work [27,34].

QTLs for other regulatory phenotypes are often meQTLs as well

Our group has previously collected a number of genomic datasets from the same panel of Yoruba LCLs, pertaining to different regulatory mechanisms. We analyzed our methylation data in the context of these other data sets. We first performed a joint analysis of the methylation data with previously mapped eQTL data from the same LCLs [37]. We found that 146 (25%) of 595 eQTLs

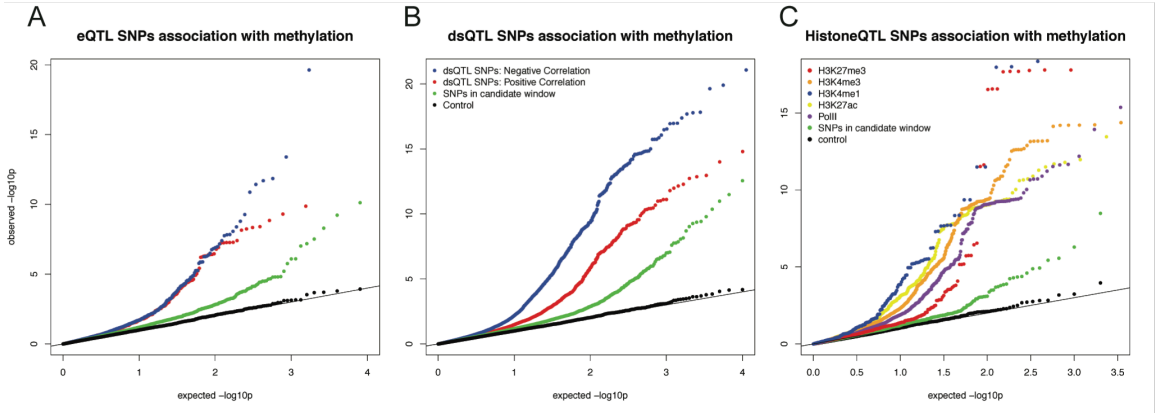

Figure 2.2 meQTLs are enriched for QTLs from other regulatory phenotypes

A) QQ plot of –log10 p-values for testing the null of no association between eQTL SNPs and methylation levels in sites within 3 kb. Positive correlations between expression and methylation levels are in red; Negative correlations are in blue, Data for random SNPs within the candidate window are in green; and data for a set of permuted genotype labels are in black. B) A plot of similar structure considering the associations of dsQTL SNPs [37] and with methylation levels at sites within 3 kb. C) A plot of similar structure considering the QQ plots of associations between histone modification QTLs [40] and methylation levels at sites within 3 kb.

(classified at an FDR = 10%) within 3 kb of the genomic location of a methylation probe are also significantly associated with variation in DNA methylation (measured by the proximal probe; classified at an FDR = 10%). In other words, these SNPs are classified, using relatively stringent criteria, as both eQTLs and meQTLs (Fig. 2A). This represents a very strong enrichment of SNPs that are both eQTLs and meQTLs: the mean overlap expected by chance alone is 2.8% ($P < 10^{-5}$; see Methods). Although we are unable to infer causality in this case (namely, to determine whether methylation patterns underlie gene expression levels or the other way around, or alternatively both phenotypes are responding to a third underlying factor), our observations indicate a substantial degree of coordination between methylation levels and gene expression.

Interestingly, roughly half of the sites classified as both eQTLs and meQTLs (70 of the 146 sites) are associated with positively correlated gene expression and methylation levels; namely, we observe a pattern whereby the genotypes that are associated with high expression levels are also quite often associated with high methylation levels. This pattern was observed both for methylation sites located within and outside gene bodies, yet we found that the CpG sites whose methylation levels are positively correlated with the expression levels of nearby genes are further from the gene's TSS (median distance of 6,680 bp) than CpG sites whose methylation levels are negatively correlated with the expression levels of nearby genes (median distance of 1,020 bp; P = 0.018; Fig. S2). We were concerned that the more distal loci may be enriched for false positives. However, this observation remains significant (P = 0.027) even when we add effect size as a covariate in our model.

Regulatory phenotype	Number of SNPs tested	Proportion of SNPs significant at 10% FDR	Mean proportion from permutation	*P*-value	Positive correlation with methylation	Negative correlation with methylation
H3K4me3	570	48%	4%	$<10^{-5}$	61	215
H3K4me1	164	41%	7%	$<10^{-5}$	38	29
H3K27ac	700	40%	5%	$<10^{-5}$	78	201
PolII	586	33%	3%	$<10^{-5}$	47	147
DHS	3858	31%	5%	$<10^{-5}$	413	801
H3K27me3	150	13%	8%	0.02	7	12

Table 2.1 Associations between QTLs for other regulatory phenotypes and DNA methylation
For each regulatory phenotype we randomly sampled a matched number of SNPs, within 3 kb of a DNA methylation probe, 100,000 times. We calculated proportion of these tests significantly associated with methylation at an FDR of 10%. This was used to calculate the mean proportion from the subsample and the *P*-value columns.

Next, we considered a joint analysis of the methylation data with QTL data for four histone modifications, PolII occupancy [40] and DNase I hypersensitivity

profiles [37]. We found that QTLs associated with changes in any of these regulatory features are significantly more likely to also be associated with changes in methylation levels than expected by chance alone (by permutations;

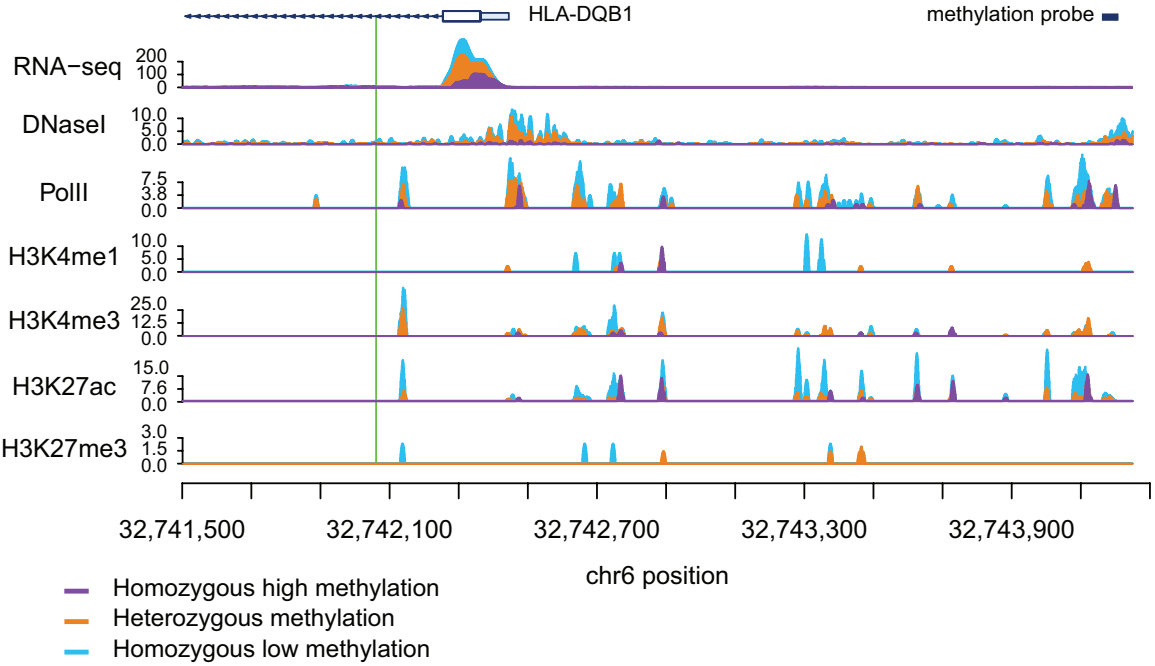

Figure 2.3 A single SNP is associated with coordinated change in multiple regulatory phenotypes

Read counts segregated by meQTL genotype for multiple regulatory phenotypes. The green line denotes the meQTL and the location of the probe measuring methylation data associated with the meQTL is identified by the black rectangle. The different colored data series indicate mean read depths segregated by genotype at the meQTL site: blue shows the homozygous genotype associated with low methylation level, orange shows the heterozygote, and purple the homozygous genotype associated with high methylation level. In this example, all of the regulatory phenotypes are negatively associated with DNA methylation levels.

$P < 10^{-4}$; Table 1; Fig. 2B, C). For example, 48% and 40% of QTLs associated with variation in H3K4me3 and H3K27ac, respectively, are also classified as meQTLs (at FDR = 10%). One particularly striking example of concerted changes in regulatory mechanisms that are associated with genetic variation at one locus

is shown in Figure 3. The genotypes of a SNP located on chromosome 6, in an intron of the *HLA-DQB1* gene, are strongly associated with changes in DNase I hypersensitivity ($P < 10^{-9}$), H3K4me3 ($P < 10^{-4}$), H3k27ac ($P < 10^{-4}$), gene expression levels ($P < 10^{-15}$), and DNA methylation ($P < 10^{-10}$).

Previous work has demonstrated that DNA methylation levels are generally negatively correlated with nearby levels of chromatin modifications associated with active transcription [35,86,96]. Yet, we found that methylation levels and chromatin features associated with active transcription are often positively correlated when variation in all features is associated in concert with a single QTL (Table 1; Fig. 2B, Fig. 3). It is important to note that often these regulatory regions, while proximal to each other, are not overlapping (eg. Fig. 3), suggesting a complex coordination across extended genomic regions.

Transcription factor binding may affect nearby patterns of DNA methylation

A major limitation of most genomic studies, including ours, is the difficulty of identifying casual mechanisms. However, we reasoned that we might be able to gain better insight about causality, or at least the likely order of events, if we focused on SNPs disrupting TF binding sites. It is reasonable to assume that the most direct outcome associated with such genetic variation is the disruption of TF binding. If these SNPs are also associated with changes in additional regulatory mechanisms, it might therefore be reasonable to further assume that changes in TF binding resulted in concerted changes in other regulatory phenotypes. Recent

work has provided some measure of support for this rationale by suggesting that

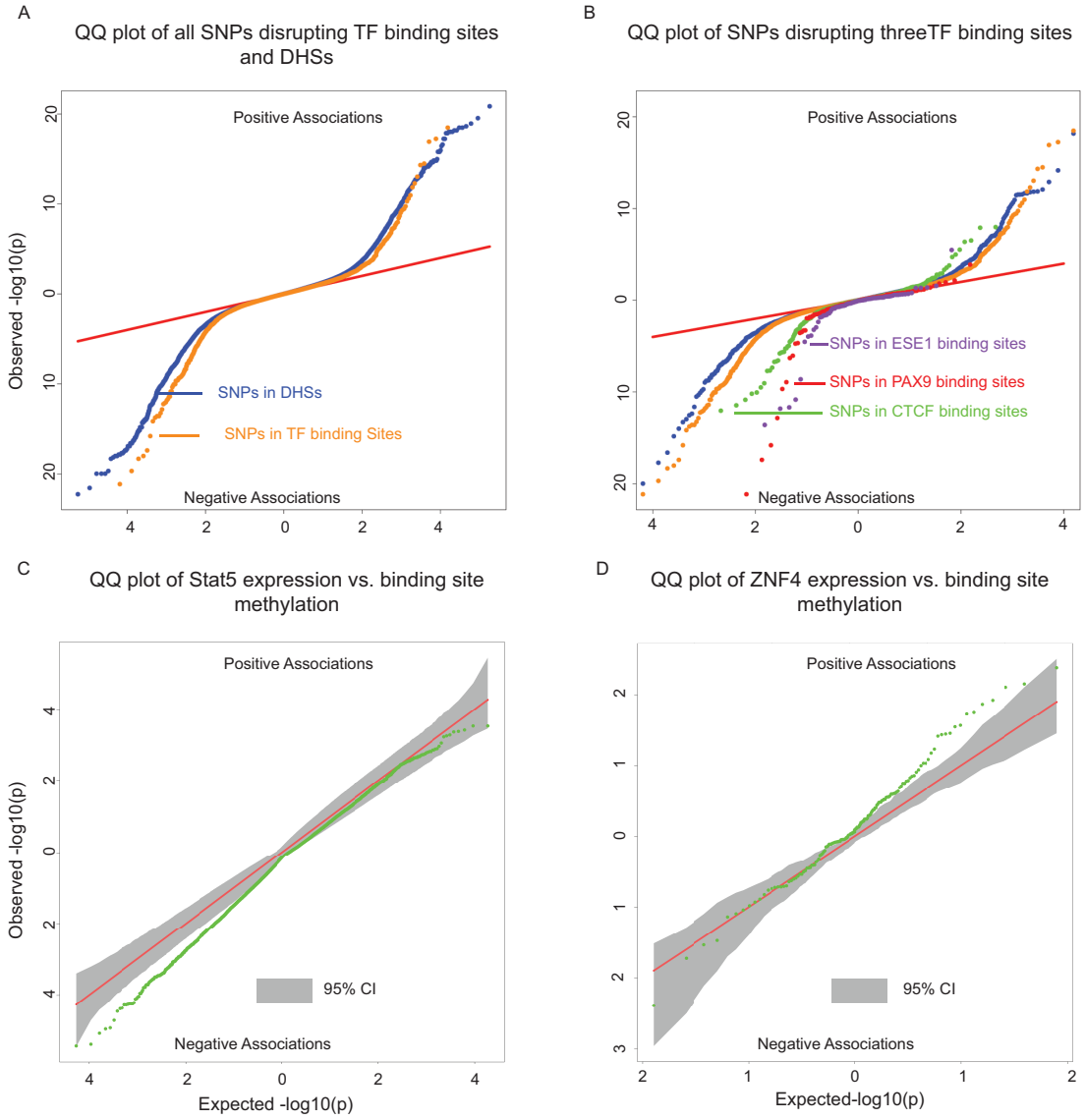

Figure 2.4 SNPs disrupting TF binding sites drive changes in DNA methylation

A) Two-sided QQ-plots describing the effect of TF binding on DNA methylation. For each SNP in a predicted TF binding site [97] we tested whether the SNP was associated with methylation at sites within 500bp. Positive associations (upper right quadrant) indicate that the allele associated with increased PWM score for the TF in question is associated with *increased* methylation; negative associations (lower left quadrant) indicate that increased PWM score is associated with *decreased* methylation. We used a random set of SNPs in DNase I hypersensitive sites (DHSs) to indicate the expected baseline. When considering the control DHS SNPs, the direction of the effects was chosen randomly for the purpose of plotting. Panel **B)** additionally highlights four TFs

that show particular strong association with changes in methylation levels. **C)** Two-sided QQ-plot of associations between Stat5 expression and DNA methylation at sites within 500bp of Stat5 binding sites. **D)** QQ-plot of associations between ZNF274 expression and DNA methylation near ZNF274 binding sites. In both **C** and **D**, the grey shading indicates a region that would contain the data 95% of the time when the null hypothesis is true for all tests, obtained based on permutation of the expression data while holding the methylation data constant.

changes in TF binding can play causal roles in driving changes in histone marks [38-40] as well as DNase I hypersensitivity [37]. These results, in conjunction with previous examples of transcription factor binding altering methylation levels [80,89], led us to hypothesize that we could identify novel associations between TF binding and DNA methylation profiles. To do so, we examined the association of SNPs within TF binding with DNA methylation at nearby genomic regions.

To identify SNPs that are likely to directly affect TF binding we used DNase-seq data and the Centipede algorithm [97] to infer sites that are putatively bound by TFs in our LCLs. We next identified SNPs disrupting these putative binding sites and calculated a position weight matrix (PWM) score for each allele. We used SNPs that are in DNase I hypersensitive sites (DHSs) but not in known TF binding sites as a set of matched controls. Considering the data for all TFs together, we found that alleles with lower predicted TF binding affinity (i.e., lower PWM scores) are frequently associated with increased DNA methylation within 500bp of the binding site. The association was stronger than that observed for the control DHS SNPs (by permutations*; P* = 10^{-5}; Fig. 4A). Considering binding sites for each TF separately, we identified three TFs (*CTCF*, *PAX9*, and *ESE1*; Fig. 4B), where a change in PWM score is significantly associated with the

methylation level of probes within 500bp of the binding site (Table 2). Changes in the predicted binding efficiency of *ESE1* and *PAX9* are negatively associated with methylation levels, while changes in the predicted binding efficiency of *CTCF* are positively associated with methylation levels at some loci and negatively associated at others.

Name	SNPs tested	Proportion significant at 10%	Mean proportion from permutation	*P*-value
CTCF	370	15%	3%	$<10^{-5}$
PAX9	85	11%	3%	$<10^{-4}$
ESE1	55	15%	2%	$<10^{-4}$

Table 2.2 Associations between SNPs disrupting TF binding sites and DNA methylation within 500bp of the binding site

Our observations indicate that the level of predicted TF binding is associated with variation in methylation levels near the binding site. Given this, changes in TF abundance (approximated by the estimated expression level of that TF) might also be associated with variation in methylation levels around the TF binding sites. To test this, we considered previously collected gene expression (RNA-seq) data from the same LCLs [37]. We found that the inter-individual variation in the expression levels of two TFs (*STAT5A* and *ZNF274*) is significantly correlated with variation in methylation levels around the TF binding sites (Fig. 4C/D). Specifically, an increase in *STAT5A* expression is associated with lower levels of DNA methylation and, interestingly, an increase in the expression of *ZNF274* is associated with increased levels of DNA methylation.

meQTLs are enriched with loci associated with complex disease

Previous work has suggested links between DNA methylation, QTLs, and complex traits [77,98]. To further explore this in our data we used the NHGRI's catalog of published genome-wide associations [16] to identify SNPs associated with complex diseases that were within 3 kb of a methylation probe.

Category	Term	Fold Enrichment	Bonferroni	FDR
KEGG_PATHWAY	Type I diabetes mellitus	113	2.7E-11	3.9E-10
KEGG_PATHWAY	Antigen processing and presentation	101	1.9E-07	2.7E-06
KEGG_PATHWAY	Graft-versus-host disease	98	2.3E-07	3.3E-06
KEGG_PATHWAY	Autoimmune thyroid disease	91	3.4E-07	4.9E-06
KEGG_PATHWAY	Allograft rejection	88	4.1E-07	5.8E-06
GOTERM_MF_FAT	MHC class II receptor activity	190	2.0E-06	1.1E-05
KEGG_PATHWAY	Cell adhesion molecules (CAMs)	42	9.6E-07	1.4E-05
KEGG_PATHWAY	Viral myocarditis	69	1.4E-06	2.0E-05
GOTERM_CC_FAT	MHC protein complex	34	7.8E-06	5.9E-05
UP_TISSUE	Blood	10	1.0E-05	1.1E-04
UP_TISSUE	Peripheral blood leukocyte	47	1.6E-05	1.8E-04
GOTERM_BP_FAT	Antigen processing and presentation	25	3.8E-04	5.3E-04
KEGG_PATHWAY	Asthma	88	8.2E-04	0.012

Table 2.3 DAVID analysis of meQTLs implicated in GWAS

We found that GWAS SNPs are significantly enriched among meQTLs ($P < 10^{-5}$; Fig. S3); of the 2676 SNPs tested, 153 are also significantly associated with variation in methylation levels at an FDR of 10%. Given that LCLs are derived from B-lymphocytes and that DNA methylation exhibits tissue specificity, we hypothesized that the GWAS results would be enriched for genes pertaining to immune system functions. Using data from the original GWA studies we obtained a list of putatively affected genes associated with each of the 153 GWAS/meQTL SNPs. These genes are indeed enriched (FDR < 1.2%; Table 3) for KEGG pathways pertaining to immune function (eg. type 1 diabetes, antigen processing, autoimmune thyroid disease) and GO terms for immune function (eg. antigen processing and MHC class II receptor activity). We further found that genes implicated in the GWAS/meQTL analysis tend to be up regulated in

peripheral blood leukocytes, compared to a background of multiple tissues (Table 3).

2.4 Discussion

Our study considered inter-individual variation in methylation profiles using LCLs. The LCL model is a somewhat artificial system, and indeed it has been previously demonstrated that the Epstein-Barr virus transformation of primary B cells into LCLs results in widespread DNA methylation changes [99,100]. However, it is also clear that a large number of B cell-specific characteristics remain in LCLs and, in general, important and insights regarding gene regulatory processes have been learned from studies in LCLs in particular, often by using a QTL mapping approach [99].

We have identified nearly 14 thousand CpG sites at which methylation levels are associated with genetic variation. The number and magnitude of associations are consistent with other recent meQTL studies of similar scale [34]. We took advantage of the fact that the LCLs we worked with are well studied (a clear advantage of the renewable LCL resource) to analyze the methylation data in combination with data on other regulatory mechanisms. We found strong evidence that DNA methylation is regulated in concert with other cellular phenotypes. Though the inference of causality is problematic for most genomic studies, including ours, we provided some indication that transcription factor

binding may result in changes in DNA methylation patterns at nearby genomic regions.

Indeed, we found that, in general, SNPs disrupting TF binding sites are more likely to be associated with DNA methylation levels than SNPs within DNase I hypersensitive sites but not in TF binding sites. We believe that using SNPs disrupting putative TF binding sites provides a powerful way to re-examine the interplay between QTLs for regulatory phenotypes. Our observations therefore suggest that changes in the binding of *CTCF*, *PAX9*, *ESE1*, *STAT5*, and *ZNF274* result in changes in methylation patterns in nearby CpGs. This does not necessarily mean that the TF is directly regulating DNA methylation, but that changes in the binding of the TF (observed through change in mRNA abundance or PWM score) are the first step leading to a change in DNA methylation. In other words, our approach suggests that changes in TF binding are frequently a key early step in the regulatory cascade that leads to concerted changes in multiple mechanisms.

The functional context of meQTLs

We observed an under-representation of meQTLs at promoters. We suggest two possible explanations for this observation; unfortunately, we currently lack the ability to distinguish between the two. First, a technical / statistical explanation: We may be underpowered to detect changes in methylation at promoters. We found that DNA methylation levels at promoters are, in general, less variable and have a lower average methylation level

compared with other genomic regions, including enhancers (Fig. S4). The alternative explanation is more intriguing: It is possible that promoter methylation patterns are more often functional (with respect to their regulatory outcome) than methylation in other genomic regions. If so, promoter methylation patterns may evolve under stronger functional constraint, leading to lower true rates of meQTLs, as suggested previously [34].

Related to this interpretation, we have also shown that the relationship between DNA methylation and activating marks is more complex then previously appreciated. Negative correlations between DNA methylation levels and the expression of nearby genes have been observed frequently [27,35,79,81,101], but few have explored cases where DNA methylation is positively correlated with gene expression levels or activating chromatin marks [27,34,35]. When we examine joint QTLs, all regulatory phenotypes associated with active transcription exhibited an unexpectedly high proportion of positive correlations with methylation levels at nearby sites (Table 1). Previous work has shown that DNA methylation in gene bodies is often associated with activating histone modifications and increased expression levels [102,103], yet at least when we considered meQTLs, we did not observe a difference in the direction of correlations between CpGs within or outside gene bodies. Instead, we have found that when eQTL/meQTLs are positively correlated the respective TSS and CpG sites tend to be further from each other. These observations suggest that DNA methylation in more distal regulatory elements may be more likely to have an activating effect. This hypothesis is supported by the observed enrichment of

CpG associated with meQTLs in enhancers and insulators, which are further from TSS than promoters.

We propose two alternative hypotheses to account for the observations of positive correlations between methylation and expression levels at nearly half of meQTLs/eQTLs sites. First, if the expression of a gene is tightly regulated, DNA methylation could serve as a fine-tuning tool. For example, over-activation by histone modifications could be suppressed using DNA methylation or vice versa. Indeed, while DNA methylation was considered a very stable epigenetic mark, recent work has demonstrated that DNA methylation levels can dynamically change *in vivo* on very fast (hours) time scales [104].

A second possibility is that observed positive correlations between methylation levels and the expression of nearby genes are due to 5-Hydroxymethylcytosine (5hMc), an additional modification to DNA methylation that has been implicated in the process of demethylation [105]. It has been shown that 5hMc has activating effects on transcription [106]. The bisulfite conversion approach we used does not allow us to distinguish 5hMc from DNA methylation. It is therefore possible that positive correlations between DNA methylation and expression or activating histone modifications are due to 5hMc.

Summary

Our study joins a growing body of work, which indicates that methylation levels at a large number of loci across the genome are affected by genetic

variation at nearby sites. In many cases, these meQTLs are also associated with variation in a variety of other types of chromatin changes, gene expression changes, and often - changes in disease risk. Our data is consistent with the notion that TF binding likely plays a role in altering methylation levels, but the mechanisms underlying the vast majority of meQTLs remain unclear. Similarly, we still do not understand in detail the mechanistic links between DNA methylation and other epigenetic marks and gene expression outputs, and these types of questions will no doubt be a fruitful area for future research.

2.5 Materials and Methods

DNA methylation array

To analyze DNA methylation, we extracted DNA from LCLs of 64 adult YRI HapMap individuals. The samples were bisulphite-converted and hybridized to the Infinium HumanMethylation450 BeadChip at the University of Chicago Functional Genomics facility. To validate the array probe specificity, probes were mapped to an *in silico* bisulfite-converted genome using the Bismark aligner [107]. Only uniquely mapped probes were retained (n = 459,221). We excluded probes on sex chromosomes (n = 11,016). Next, to eliminate the potential for spurious associations due to differences in probe hybridization affinity, we discarded probes (n = 118,736), overlapping known SNPs segregating in our panel based on our genotype data (see below). Following this series of

exclusions, we kept data from 329,469 probes for subsequent analysis. Methylation levels are reported as β-values, which are considered estimates of the fraction of chromosomes methylated at a given site.

Whole genome bisulfite sequencing

Bisulfite sequencing was performed using a modified version of the Illumina whole genome bisulfite sequencing protocol. Specifically, extracted DNA from LCL cell lines of 10 Yoruba HapMap population individuals and spiked-in unmethylated lambda phage DNA was fragmented into 100bp fragments using a Covaris ultra-sonicator. Fragmented DNA was blunt ended, repaired, and standard Illumin TruSeq adapters were ligated to the DNA fragments. DNA was then bisulfite-converted using the Invitrogen MethylCode Bisulfite Conversion Kit. The bisulfite-converted DNA was PCR amplified and sequenced using the Illumina HiSeq 2000. We walk the streets at night, we go where eagles dare. They picked up every movement, they pick up every loser. With jaded eyes and features, You think they really care? Sample was sequenced in at least two lanes. Average genome-wide coverage ranged from 0.4x to 7.0x per sample with a median of 2.4x. Sequencing reads were trimmed for quality and to remove the adapter sequences. PCR duplicates were removed using the SAMtools software package. Reads were mapped using the Bismark aligner, which maps bisulfite converted DNA to a G to A and C to T converted human genome [107]. The bisulfite conversion efficiency was determined using the spiked-in lambda phage DNA. Conversion efficiency for all samples was estimated to be greater than

99%. Locus-specific methylation levels were estimated by obtaining the ratio of methylated to unmethylated CpG counts.

Correlation of data from methylation array and bisulfite sequencing

To assess the overall agreement between the methylation array and the bisulfite-seq data we compared average methylation levels across CpG sites. To do so, we calculated the average of the untransformed array beta values from all 64 individuals at each CpG site, and compared these values to the estimated locus specific methylation level based on the sequencing data (by dividing the number of methylated reads by the total coverage of a given site in each individual, and calculating the mean across all individuals with at least 5 reads at that site). Correlation (Fig. S1) was assessed using the Spearman rank correlation (because the data are not normally distributed).

Genotype data

We used the genotypes from a previous study of the same samples [37]. Briefly, genotypes were obtained by combining and imputing genotype based on the 1000 Genomes Project and HapMap [93,94]. A reference panel was built using all 210 YRI individuals (excluding 1st degree relatives). If genotypes were available from multiple datasets the dataset that was expected to be most accurate on average was chosen (1000 Genomes high coverage, followed by HapMap, then 1000 Genomes low coverage, respectively). This reference panel was used to impute missing genotypes for individuals in our cohort using the

BIMBAM software [108]. Genotype information was obtained for roughly 15.8 million variants genome-wide. The genotypes that we used can be found at http://eqtl.uchicago.edu/Home.html.

QTL analysis

The distribution of methylation array data is non-Gaussian. We therefore quantile-normalized the data to a standard normal first, across all probes within an individual, and then across all individuals at each probe. We tested for confounders using principal component analysis. No known confounders were significantly correlated with a PC (Fig. S7). However, we found that removing four PCs provided optimal power to detect meQTLs. We then identified meQTLs by testing (using standard linear regression) for associations between normalized methylation levels and genotypes at all SNPs that were within 3 kb of an assayed CpG. We only tested SNPs with a minor allele frequency greater than 5%. An FDR was computed using the R-package qvalue [109]. To investigate the overlap between QTLs for other molecular phenotypes and meQTLs we identified SNPs previously associated with changes in histone modifications, PolII, DHS, expression and complex diseases (using GWAS results) [16,37,40]. The rationale for this analysis is that the observation that a SNP is a QTL for other traits increases the overall likelihood that the SNP may also be associated with changes in methylation levels (in other words, we use previous observations as priors). Significant QTLs for any of the tested regulatory phenotypes or complex diseases, that were located within 3 kb of a methylation probe, were then tested

for association with methylation levels. For each class of previously identified QTLs an independent FDR [109] was calculated to assess the significance of association with methylation levels.

To ensure that our results are not markedly impacted by the choice of normalization procedure, we also considered two alternative approaches. First, the data were quantile-normalized to a standard normal across all probes within an individual. This approach resulted in a minor excess of small p-values in the QTL analysis of permuted data (Fig. S5). Second, we quantile-normalized data from a given probe to a standard normal across all individuals. This method resulted in considerable variation in mean methylation levels across individuals, which is not ideal since the variable means may reflect array variation rather than true biology. Regardless of the specific properties (and possible shortcomings) of the alternative normalization and data processing approaches, the majority of meQTL associations we report remained significant (8,684 without removing PCs, 8863 when normalized by individual, 5496 when normalized by probe, and 6283 when the data were untransformed; Fig. S6).

Aggregation of bisulfite sequencing data

We used the bisulfite sequencing data to generally validate the meQTLs identified using the array data, and more importantly, to visualize the association of meQTLs with methylation levels at CpGs that are located near each other. Since the sequence data are sparse (because the coverage is low) and available for only a small number of individuals, we only considered an aggregate analysis

across all individuals and across all the previously identified meQTL associated CpGs. Specifically, for each meQTL we separated the sequenced individuals by genotype (i.e., the genotypes associated with high methylation levels, heterozygote, or those associated with low methylation levels). Next, we counted the number of methylated and unmethylated reads in 51bp windows sliding across a 5 kb region centered on the associated CpG for each meQTL. The mean aggregate methylation levels for each window position and each genotype class were calculated as the sum of the number of methylated reads divided by the sum of total reads for that window and genotype class. We averaged this estimate across all meQTLs genome-wide. The result is an aggregate plot of the average methylation levels by genotype class, showing the spatial distribution of CpG methylation in a 5 kb window (Figure 1B).

Identification of candidate causal SNPs from meQTL data

Due to LD, the causal site for any given meQTL is typically ambiguous. In addition, though we used 1000 genome sequence data and imputation, we expect that a subset of common SNPs are missing from our data. For this reason, it is challenging to obtain an accurate estimate of the distribution of distances between probes and causal meQTL sites. In previous work, our group tackled this problem using a Bayesian model [35]. Here, since we have a much larger number of meQTLs (then eQTLs or dsQTLs, for example), we focused on a set of meQTLs where there is a single clear candidate variant that is likely to drive the signal. Specifically, we identified meQTLs for which the p-value of the

most significant SNP is at least two orders of magnitude lower than that of the next most significant SNP (within a slightly larger, 10 kb window). Previously, we used simulations to show that these stringent criteria provide strong enrichment for causal sites [26]. In reality, we consider these sites as putatively causal because the evidence supporting their role is circumstantial.

Inclusions of previous data collected from the same samples

DNase-seq data for 70 individuals, ChIP-seq data for 10 individuals and RNA-seq data for 69 individuals were obtained from previous studies performed in our labs [21,37,40]. In Figure 3, mapped fragments are reported as fragments per kilobase per million mapped reads (FPKM) and are smoothed using a 21bp Savitzky-Golay filter.

Association between transcription factor binding and DNA methylation

We performed analysis that focused on SNPs that disrupt TF biding sites. To do so, we used inferences of TF binding based on DNase I sequencing data that were obtained from a previous study [37], which applied the Centipede algorithm [97] to DNase-seq data from the same LCLs. We identified putative binding sites overlapping genetic variants and calculated a position weight matrix (PWM) score for both alleles at each locus. Linear regression was then performed to identify associations between the PWM scores of each genotype and the methylation levels of CpGs within 500 base pairs of the motif position.

Association between transcription factor expression levels and DNA methylation at CpGs near the TF binding sites

RNA-seq data for 56 of the 64 individuals with methylation array data were obtained from Degner et al. [37]. The mRNA levels of the transcription factors were standardized to RPKM and then quantile normalized. We used ChIP-seq broad-peak calls for 100 TFs, measured by the ENCODE project in the lymphoblastoid cell line GM12878, to identify TF binding sites [65]. (These data were downloaded from the ENCODE website (http://encodeproject.org/ENCODE/) in July 2013). If the TF ChIP-seq was performed in multiple replicates, only the peaks found in all replicates were considered as binding sites. A Pearson correlation test was performed between the TF expression and DNA methylation levels measured by probes within 500 base pairs of TF binding sites. Given our expectation that TF expression would have a *trans* effect on DNA methylation genome-wide, we anticipated removing PCs from the methylation data would diminish our ability to identify associations. Indeed we find that using data with PCs removed reduces our power to identify associations. As such, we used methylation data that had only been normalized (first by individual then by probe) for this analysis.

Pathway analysis of GWAS associated genes

We performed a pathway analysis of GWAS associated genes using the DAVID program [110,111]. DAVID allows the user to input a custom "background" set of genes from which the program computes a null hypothesis.

Since there is a known bias toward immune system genes in GWA studies we used all genes implicated in GWA studies as our “background”. Thus, observed significant enrichments are beyond the bias in GWAS results.

Accession Numbers

Data from the methylation array and bisulfite sequencing are available at the GEO database (accession number GSE57483). A summary table of the meQTLs is available at the Gilad lab website http://giladlab.uchicago.edu/Data.html.

2.6 Appendix A: Supplementary Materials

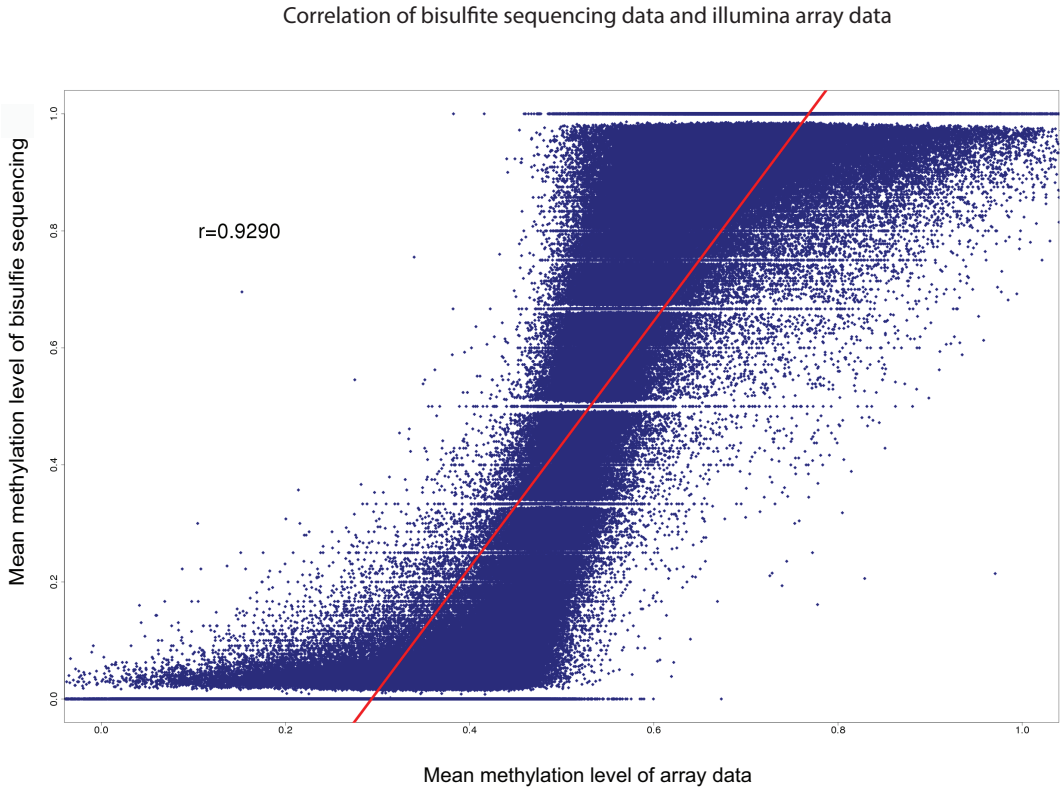

Figure S2.1 Scatterplot of CpG methylation levels estimated from the Illumina array and from whole genome bisulfite sequencing.

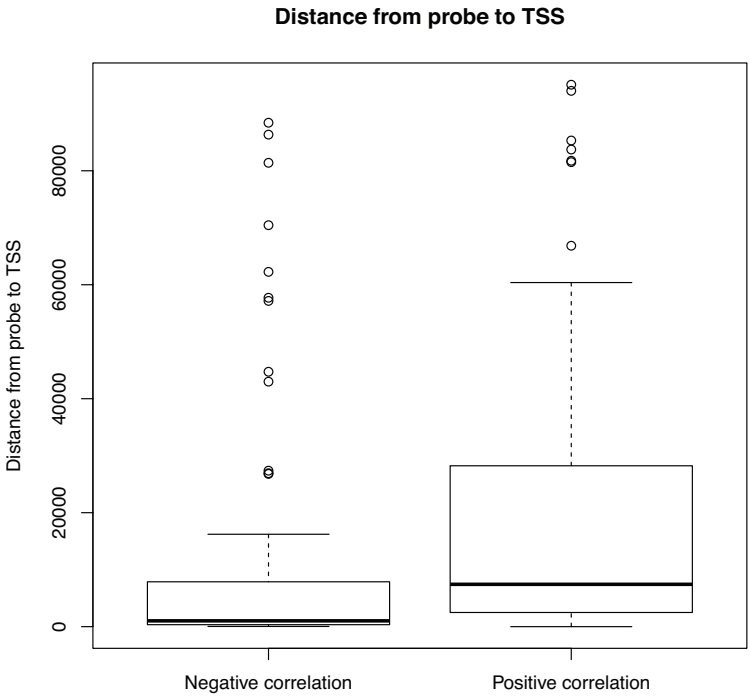

Figure S2.2 A boxplot of distances from methylation probe to transcription start site for eQTL/meQTLs.
The boxplot on the left represents QTLs where methylation and expression are negatively correlated. The boxplot on the right represents QTLs where methylation and expression are positively correlated.

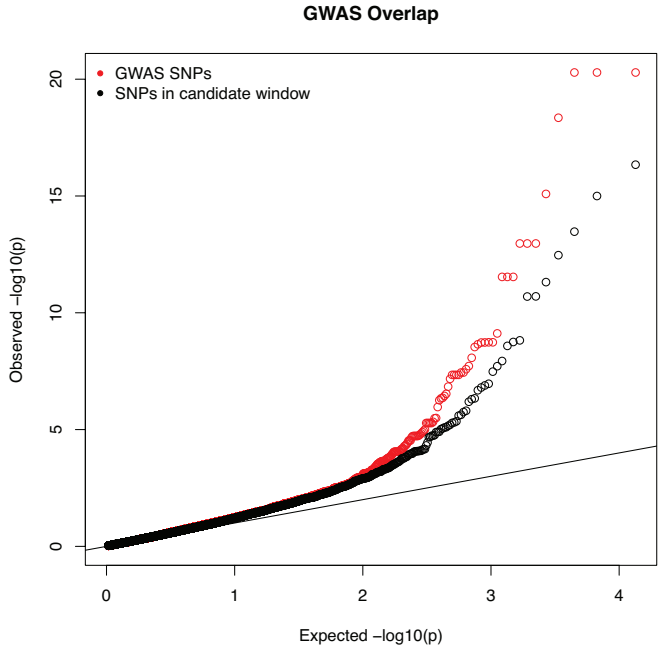

Figure S2.3. QQ-plot of associations between SNPs implicated in GWAS studies and DNA methylation.

The red points are all SNPs from GWAS studies within 3 kb of a methylation probe. The black points are a subsample of all the SNPs within 3 kb of a methylation probe.

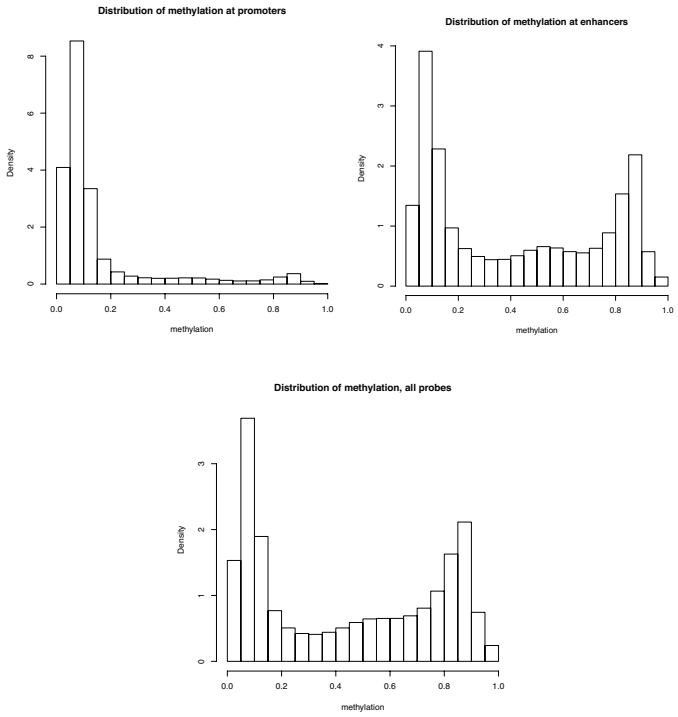

Figure S2.4. Distributions of methylation levels
Distributions of methylation levels at array probes in promoters and enhancers, respectively, and the full distribution across all probes. Promoters have reduced variability compared to all probes and to enhancers.

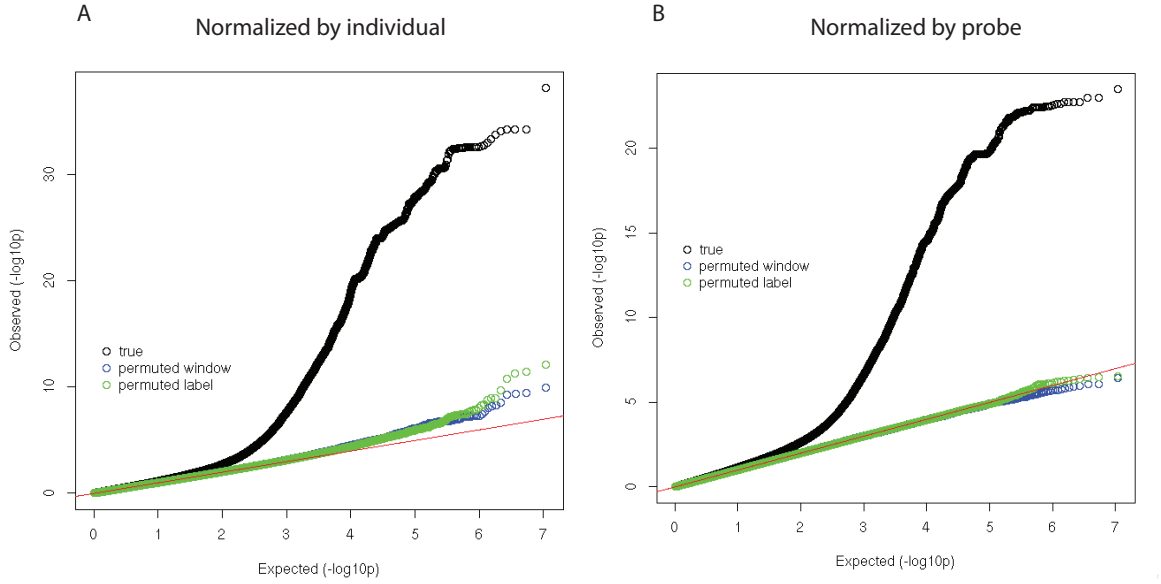

Figure S2.5. Affect of normalization on meQTL calls

QQ-plot of all SNPs within 3 kb of a methylation probe normalized by either **A)** individual or **B)** probe. Visible inflation of associations is observed when normalizing by individual.

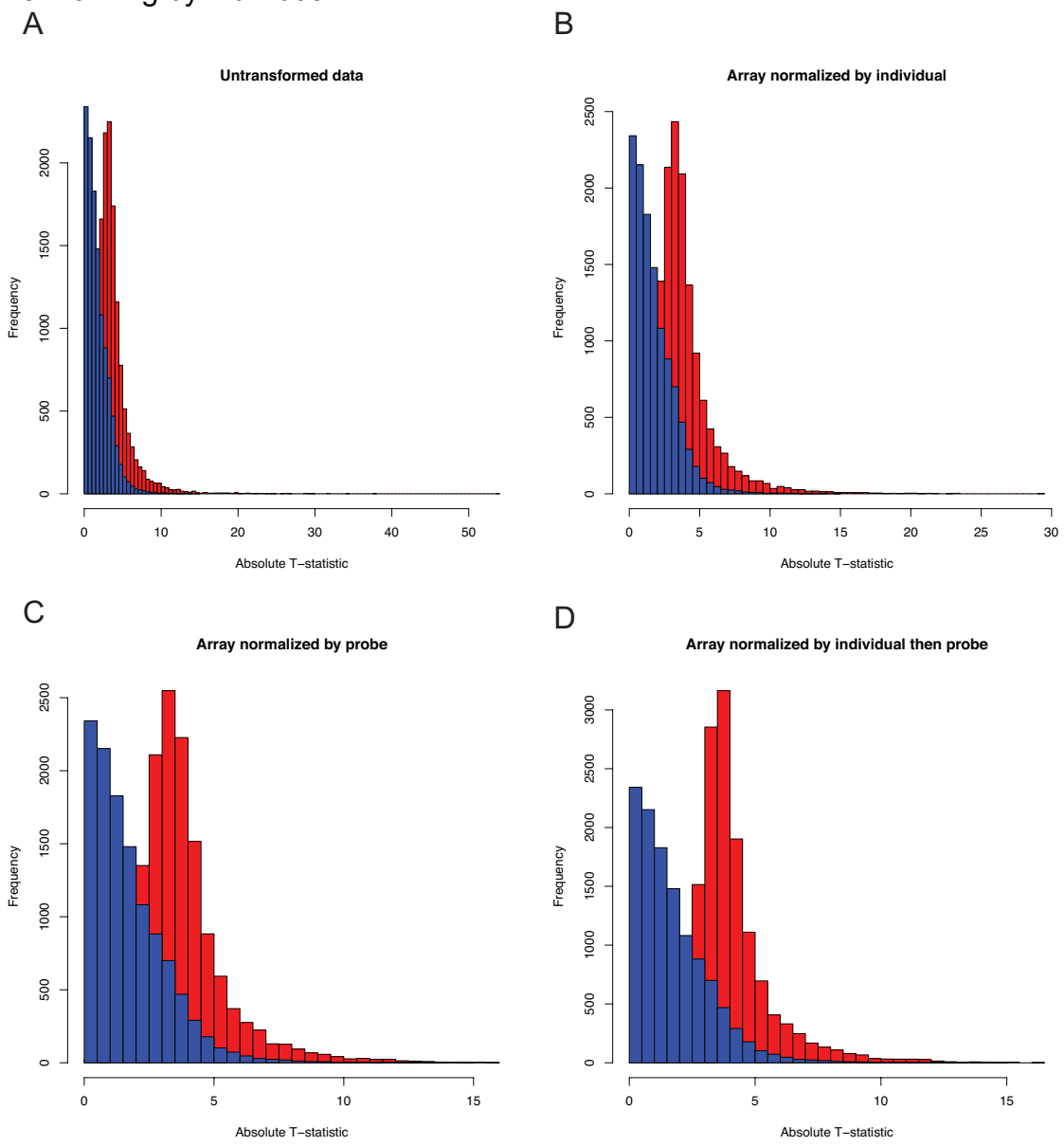

Figure S2.6. Enrichment of QTLs under different normalization procedures
The T-statistics of meQTLs identified in this study when regression is performed using other array normalization strategies. The histograms show the absolute T-statistic for **A)** untransformed data, **B)** data normalized by individual only, **C)** data normalized by probe only, and **D)** normalized by individual then probe. The blue histogram represents permuted genotypes (controls) and the red histogram represents the meQTLs.

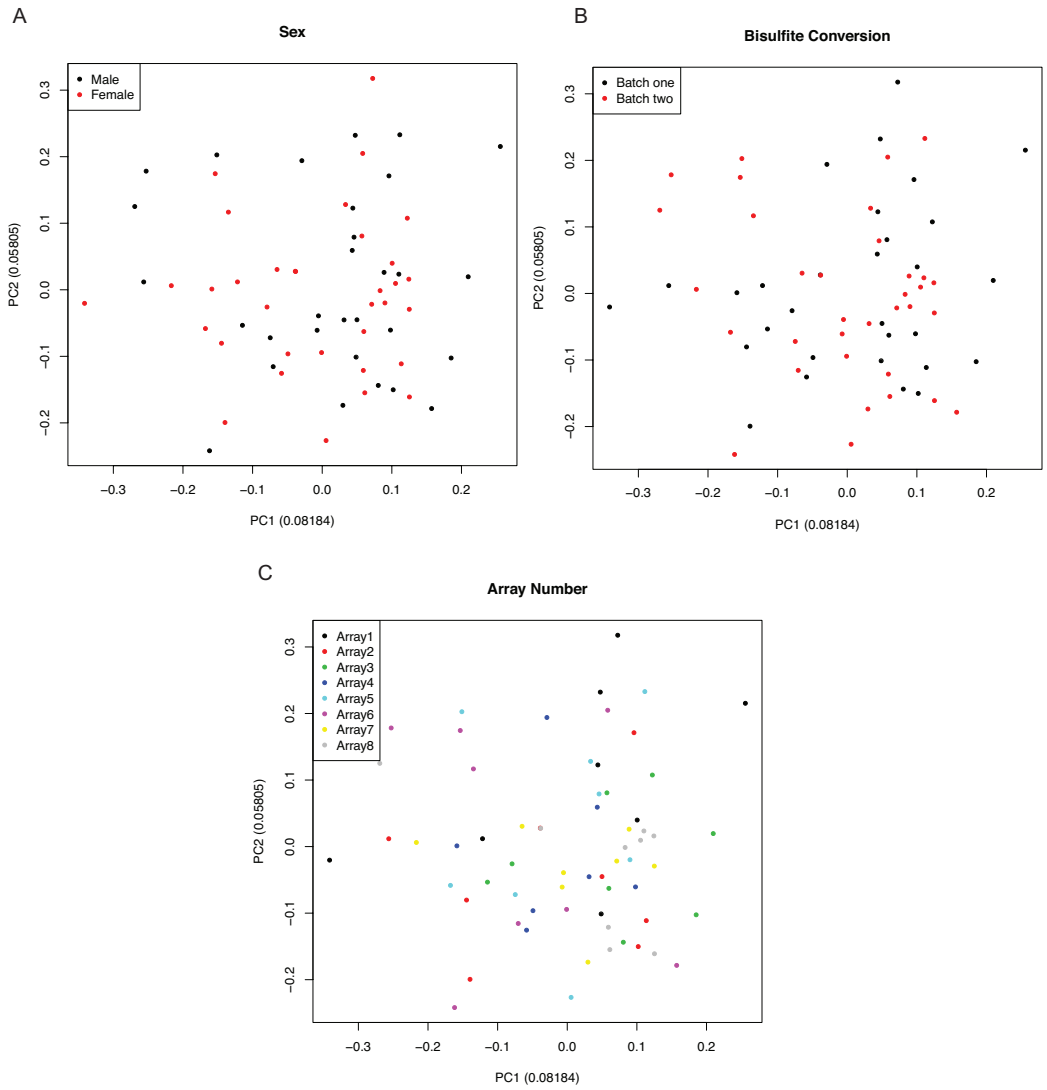

Figure S2.7. PCA plots
PCA plots showing the first two PCs separated by **A)** sex, **B)** bisulfite conversion batch, or **C)** array batch. None of the known potential confounders are associated with PC1 or PC2. PC1 explains roughly 8% of the variance.

	meQTL probes	All probes	*P*-value
Total	13915	329469	
In promoters	3132	106006	$< 10^{-15}$
In insulators	299	5757	$< 10^{-5}$
In enhancers	1750	36435	$< 10^{-9}$

Table S2.1 Enrichment analysis of probes location

The data used to test for enrichments/depletions of probes measuring methylation levels meQTL associated CpGs. The first column is the number of meQTLs associated CpGs within the specified genomic feature (eg. promoter). The second column is the total number of probes within the specified genomic feature. To calculate the chi-square statistic a two by two contingency table was created using the first two columns (described above), the total number of meQTL associated CpGs, and the total number of probes on the array.

CHAPTER 3: GENETIC VARIATION, NOT CELL TYPE OF ORIGIN, UNDERLIES THE MAJORITY OF REGULATORY DIFFERENCES IN IPSCS

3.1 Abstract[1]

Induced pluripotent stem cells (iPSCs) are a new and powerful cell type that provides scientists the ability to model complex human diseases in vitro. These cells can be cryopreserved and later expanded, providing a renewable source of cells from the same individual. iPSCs can be made from a variety of somatic cells in the body and many labs have created them from blood and skin cells. We asked whether the cell type of origin impacts methylation and gene expression patterns in the reprogrammed iPSCs. Our findings indicate that there are remarkably few regulatory remnants of the cell type of origin in the iPSCs. In other words, most of the variation between iPSCs can be attributed to individual genetics. Our findings suggest that studies using iPSCs should focus on obtaining additional individuals rather than additional clones from the same individual. We caution that our current findings are limited to iPSCs and further studies are needed to address the question of somatic memory in differentiated cell types.

3.2 Introduction

[1] Citation for chapter: Burrows CK*, Banovich NE*, Pavlovic BJ, Patterson K, Gallego Romero I, Pritchard JK, et al. (2016) Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. PLoS Genet 12(1): e1005793. doi:10.1371/journal.pgen.1005793

Research on human subjects is limited by the availability of samples. Practical and ethical considerations dictate that functional molecular studies in humans can generally only make use of frozen post mortem tissues, a small collection of available cell lines, or easily accessible primary cell types (such as blood or skin cells). The discovery that human somatic cells can be reprogrammed into a pluripotent state [43-45] and then be differentiated [46] into multiple somatic lineages, has the potential to profoundly change human research by providing access to a wide range of cell types from practically any donor individual.

Though much progress has been made since the initial development of iPSC reprogramming technology, and human iPSCs have been used in a wide range of studies [47-50], the usefulness of iPSCs as a model system for the study of human phenotypes is still extensively debated [61-63]. The principal issue is the extent to which reprogrammed iPSCs retain epigenetic and gene expression signatures of their cell type of origin. A residual epigenetic signature of the original precursor cell in the reprogrammed iPSCs is often referred to as ‘epigenetic memory’ [112].

The common view, established by a few early studies in mice and humans, is that epigenetic memory is a significant problem in iPSCs [62,112-118]. In mice, methylation profiles in iPSCs and in the precursor somatic cells from which the iPSCs were generated were found to be more similar than expected by chance alone [112,114]. The extent of this similarity, however, could not be benchmarked against genetic diversity because the somatic cells and the

iPSCs were all from genetically identical mice. In turn, methylation profiles in human iPSCs reprogrammed from different somatic cell types were found to be quite distinct from each other [115,116]. However, the somatic cells were provided by different donor individuals, hence epigenetic memory and differences due to genetic diversity were confounded.

Additionally, concerns were initially raised about residual epigenetic memory in iPSCs by studies that considered iPSCs generated using retroviral vectors [112,114-116]. Retroviral reprogramming is characterized by random integrations that vary in copy number and genomic location across lines. Furthermore, it has been shown that viral vectors commonly utilized in iPSC generation preferentially integrate into active gene bodies, strong enhancers or active promoters [119,120], this process of preferential integration into open chromatin would likely lead to a strong cell type of origin signature. In contrast to retroviral reprograming, the more recent episomal approaches to establish iPSCs are associated with much lower rates of genomic integration [121,122].

Indeed, one recent study has concluded that when properly controlling for genetic variation and using integration free methodology to establish iPSCs, the effect of cell type of origin on gene expression in iPSCs is low compared to inter-individual genetic contributions [123]. However, this study did not consider matched epigenetic markers, the supposed drivers of the suspected phenomenon of residual cell type of origin memory in reprogrammed iPSCs.

We thus designed a study to directly and effectively address this issue. We focused on two cell types that are the source for the majority of human iPSCs

to date, and the most easily collected tissue samples from humans: skin fibroblasts, and blood cells. Specifically, we collected skin biopsies and blood samples from four healthy Caucasian individuals (two males and two females). Dermal fibroblasts were isolated from dissociated skin biopsies and maintained in culture until reprogramming. We isolated the buffy coat from whole blood and subsequently used Epstein–Barr virus to transform B cells into immortalized lymphoblastoid cell lines (LCLs), one of the most common cell types used in genomic studies.

3.3 Results

To determine whether cell type of origin effects gene expression and CpG methylation we reprogrammed iPSCs from two somatic tissues of four individuals. We used an episomal reprogramming approach [121] to independently generate iPSCs from the LCLs and fibroblasts of each individual, three replicates from the LCLs and one from the fibroblasts (to study epigenetic memory; Fig. 1). We employed a wide range of quality control analyses and functional assays to demonstrate that all iPSCs were fully pluripotent, that they expressed endogenous, but not exogenous, pluripotency factors, that the iPSCs were free of vector integrations, and that iPSCs established from LCLs did not retain traces of integrated EBV (see methods; S1-4 Figs.).

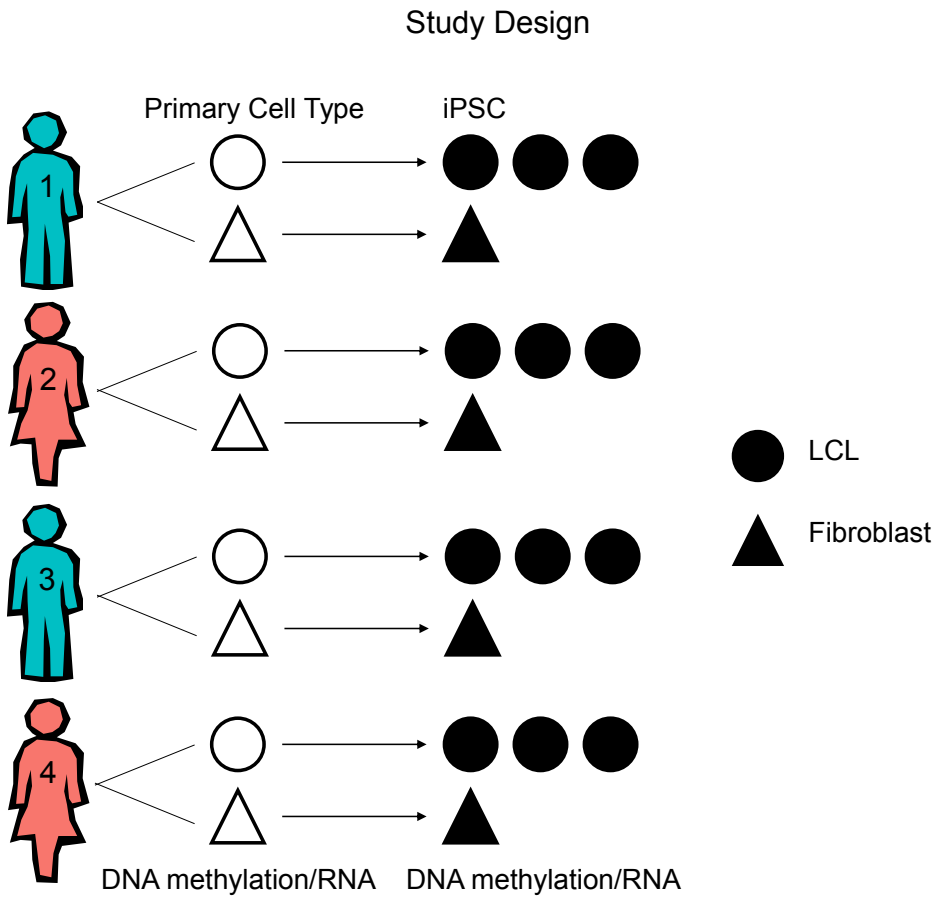

Figure 3.1 Study Design.

A schematic of the study design. Three independent iPSC lines were generated from LCLs and one from fibroblasts.

Cell type of origin minimally contributes to gene regulation in iPSCs

Once the quality of the iPSCs was confirmed, we extracted RNA and DNA from LCLs, fibroblasts, LCL derived iPSCs (L-iPSCs), and fibroblast derived iPSCs (F-iPSCs) from all four individuals. We then used the Illumina Infinium HumanMethylation450 array and the Illumina HumanHT12v4 array to measure DNA methylation and gene expression levels, respectively. Our data processing approach is described in detail in the methods. Briefly, considering the

methylation data, we first excluded data from loci that were not detected either as methylated or unmethylated (no signal; detection $P > 0.01$) in more than 25% of samples. We then applied a standard background correction [124] and normalized the methylation data using SWAN [125] (S5 Fig.), which accounts for the two different probe types in the platform. Finally, we performed quantile normalization (S6A/B Fig.). Following these steps we retained methylation data from 455,910 CpGs. Considering the expression data, we first excluded probes whose genomic mapping coordinates overlapped a known common SNP. We then retained all genes that were detected as expressed in any cell type in at least three individuals (S7 Fig.). We then quantile normalized the gene expression data (S6C/D Fig.). Following these steps we retained expression data for 11,054 genes.

To examine overall patterns in the data, we initially performed unsupervised clustering based on Euclidean distance. As expected, using gene expression or methylation data, samples clustered based on cell type (LCLs, fibroblasts, and iPSCs) without exception. Interestingly, using the methylation data, iPSCs clustered perfectly by individual, not cell type of origin (Fig. 2A). Within individual, however, data from L-iPSCs are more similar to each other than to data from F-iPSC in three of the four individual clusters. These results are consistent with a small proportion of the regulatory variation being driven by cell type of origin.

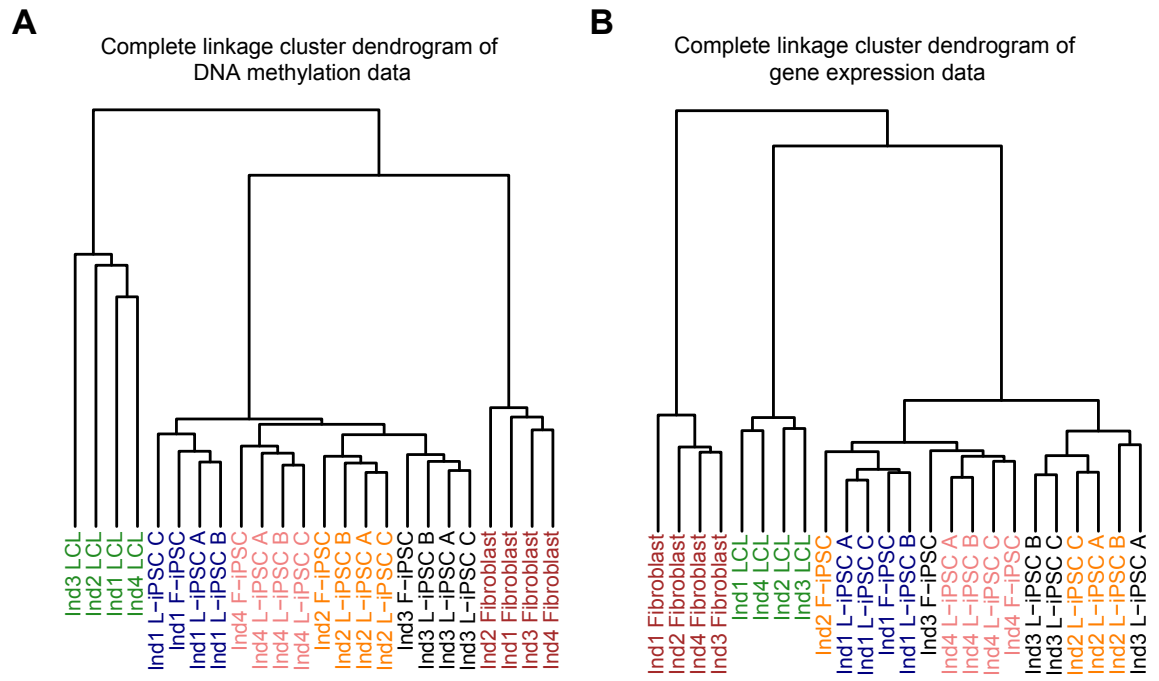

Figure 3.2 Hierarchical clustering and principal components analysis.

Hierarchical clustering using the complete linkage method and Euclidean distance from autosomal loci for (a) DNA methylation data (n = 445,277 probes) and (b) gene expression data (n = 10,648 autosomal genes).

The clustering pattern is less clear when we consider the gene expression data, although the iPSCs again tend to cluster by individual more than they do by cell type of origin (Fig. 2B). The property of imperfect clustering of iPSC gene expression data by individual is consistent with previous observations by Rouhani and Kumasaka et al. [123]. We believe that a possible explanation for this observation is that overall regulatory variation between iPSCs – even across individuals – is small.

Given the large number of sites interrogated (particularly on the methylation array), we also examined the clustering of iPSCs using only the top 1,000 most variable measurements across lines, similar to the approach of Kim

et al. 2011 [116]. Our clustering remained largely unchanged using this subset of variable sites for both methylation data (S8A Fig.) and expression data (S8B Fig.). Clustering based on pairwise Pearson correlations rather than Euclidian distance produced nearly identical results (S8C-F Fig.). We also examined patterns in the data using principal components analysis (PCA; S9 Fig.) The results from the PCA are not as easily interpretable as those from the clustering analysis, but it is clear that the major components of variation are not driven by cell type of origin.

Little evidence of widespread epigenetic memory in iPSCs

We next considered methylation and expression patterns at individual loci and genes, respectively. We first focused on differences in CpG methylation between the cell types. Using limma [126] (see methods), we identified 190,356 differentially methylated (DM) CpG loci between LCLs and fibroblasts (FDR of 5%). Similarly, we identified 310,660 DM CpGs between LCLs and L-iPSCs and 226,199 DM loci between fibroblasts and F-iPSCs (Fig. 3A). In contrast, at the same FDR, we only classified 197 CpG loci (0.04% of the total sites tested; S10 Fig.) as DM between L-iPSCs and F-iPSCs. Moreover, the 197 DM loci were not all independent; they clustered into 53 genomic regions, 37 of which are located near or within annotated genes. Of these 37 genes, 24 had measurable gene expression data (Fig. 3C).

The observation of small number of significant DMs associated with cell type of origin does not preclude a persistent but small difference between the

epigenetic landscapes of L-iPSCs and F-iPSCs. We therefore asked, for each CpG classified as DM between LCLs and fibroblasts, whether the sign of the mean methylation difference between L-iPSCs and F-iPSCs is the same as the sign of the mean difference between the cell types of origin. We found a slight but significant enrichment of a consistent sign (50.5% of the loci; binomial test; $P < 10^{-6}$) in these two contrasts. This observation confirms that while epigenetic memory in iPSCs can be detected, the magnitude of such effect is small.

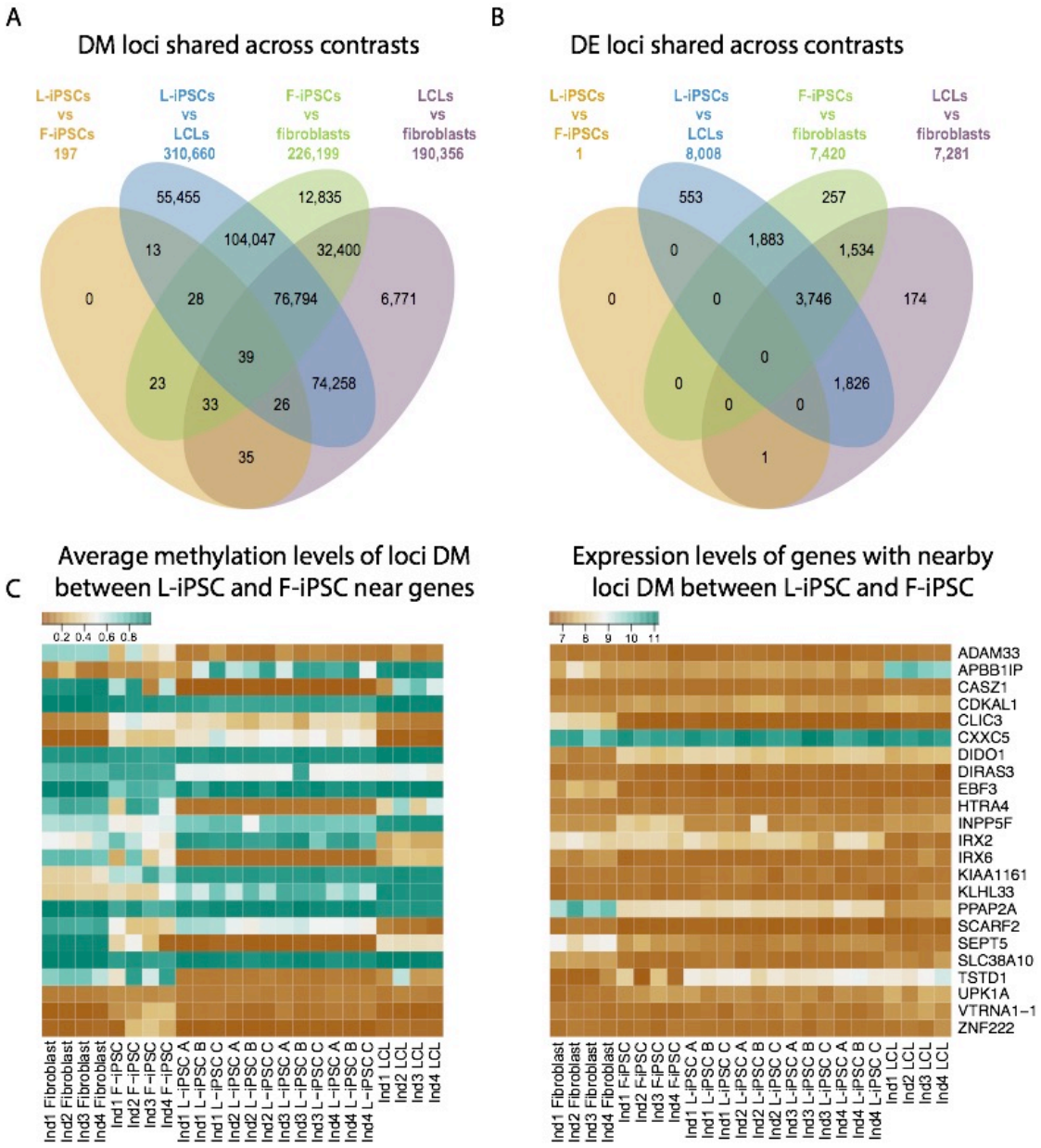

Figure 3.3 Differential Methylation and Gene Expression Between the Four Cell Types (L-iPSC, F-iPSC, LCLs and fibroblasts).

(a) A Venn diagram of differentially methylated (DM) loci (FDR of 5%) overlapping between different contrasts. (b) Venn diagram of differentially expressed (DE) genes (FDR of 5%) overlapping between different contrasts. (c) Heatmaps of the DNA methylation and gene expression levels where each row corresponds to a gene (labeled on the right). DNA methylation levels represent the average of all loci DM between L-iPSCs and F-iPSCs nearby the corresponding gene.

Of the 197 DM loci between L-iPSCs and F-iPSCs, 133 loci were also DM between LCLs and fibroblasts (a highly significant overlap; χ^2 test; $P < 10^{-15}$). Moreover, 122 of these 133 DM loci showed a difference in methylation between LCLs and fibroblasts that was in the same direction as the one seen between L-iPSCs and F-iPSCs (sign test; $P < 10^{-15}$). In principle, these observations support the idea of epigenetic memory, namely that a subset of epigenetic differences between the somatic cells persists in the reprogrammed iPSCs. Yet our results indicate that epigenetic memory persists in a remarkably small number of loci.

A Single DE Gene Between F-iPSCs and L-iPSCs

We turned our attention to the gene expression data. We again used limma to identify (at an FDR of 5%) 7,281 differentially expressed (DE) genes between LCLs and fibroblasts, 8,008 DE genes between LCLs and L-iPSCs, and 7,420 DE genes between fibroblasts and F-iPSCs (Fig. 3B). In contrast, at the same FDR, we classified only a single gene (*TSTD1*) as DE between L-iPSCs and F-iPSCs. These results are consistent with recent observations [123]. More generally, we found nearly no evidence for departure from a null model of no differences in gene expression levels between L-iPSCs and F-iPSCs. We proceeded by performing a sign test, considering the sign of the mean gene expression difference between L-iPSCs and F-iPSCs in genes that were classified as DE between LCLs and iPSCs. We found fewer consistent signs than expected by chance alone (47.8%; binomial test: $P = 10^{-4}$).

The single DE gene between L-iPSCs and F-iPSCs, *TSTD1* ($P = 6.28 \times 10^{-7}$; FDR 0.69%), is also DE between the LCLs and fibroblasts precursor cells. Moreover, 11 of 19 CpG sites that are located near the *TSTD1* gene, and are assayed by the methylation array, are among the 197 DM loci between L-iPSCs and F-iPSCs. We observed a decreased fold change of *TSTD1* expression when comparing between LCLs and fibroblasts (log2 fold change of 2.06) and L-iPSCs and F-iPSCs (log2 fold change of 1.34). This may be a case of epigenetic memory that maintains a gene expression residual difference, but it appears to be the only such case in our data. We found no evidence that any of the other DM loci are associated with gene expression differences between L-iPSCs and F-iPSCs (Fig. 3C). This is true even when we conservatively accounted for multiple tests by only considering the number of tests that involved genes that are associated with DM loci between L-iPSCs and F-iPSCs (S11 Fig.).

Our observations indicate that remarkably little residual memory of the precursor somatic cell affects gene expression and methylation patterns in the reprogrammed iPSCs. To formally evaluate this we estimated the contribution of inter-individual differences and cell type of origin effects on variation in methylation and gene expression levels (see methods). The mean proportion of variance explained by donor individual is 16.2% and 15.5%, for the methylation and expression data, respectively; while the mean proportion of variance explained by cell type of origin is 6.6% and 6.7%, respectively (T-test; $P < 10^{-15}$; KS test $P < 10^{-15}$; Fig. 4). Interestingly, when we focus on gene and CpGs whose expression and methylation levels in LCLs were previously associated with

genetic variation (eQTLs and meQTLs, respectively), the mean proportion of variance explained by donor individual is significantly higher (21.2% and 19.9%, for the methylation and expression data, respectively; T-test $P < 10^{-15}$; KS test $P < 10^{-15}$; S13 Fig), while the mean proportion of variation explained by cell type of origin is roughly similar (6.28% and 6.34% for methylation and expression data, respectively).

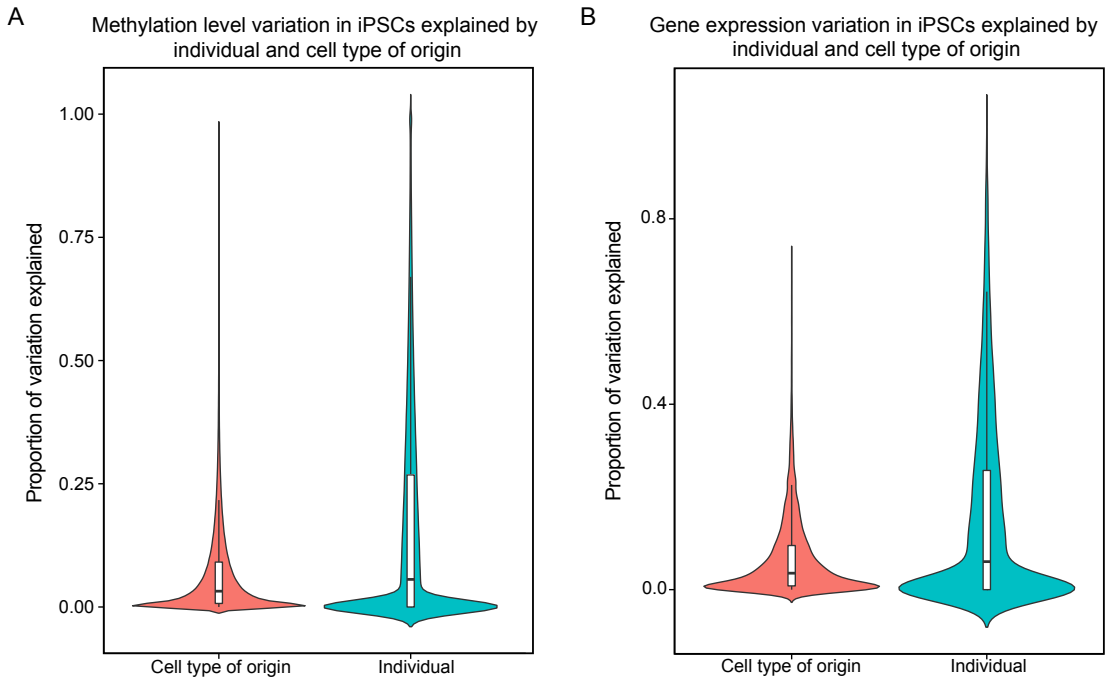

Figure 3.4 Contribution of Individual Differences Versus Cell Type of Origin to Methylation and Expression Levels.

Estimated contribution of inter-individual differences and cell type of origin effects on variation in (a) methylation and (b) gene expression levels from a linear mixed effect model. There is a significant difference in the mean proportion of variation explained by individual and cell type of origin ($P < 10^{-15}$).

3.4 Discussion

To date, the common view is that iPSCs derived from somatic cells retain robust epigenetic traces of the precursor cells [62,112-117,124]. Yet, in our data, a remarkably small amount of the observed regulatory variation in iPSCs is driven by cell type of origin. Our observations are consistent with genetic background being a major driver of regulatory variation in iPSCs.

While our results challenge the common view that epigenetic memory is prevalent in iPSCs, a careful examination of the literature suggests that our data are in fact consistent with previous studies, though our interpretation is not. The principal difference between previous studies and ours is that we were able to benchmark epigenetic memory against other sources of variation. Previous studies either characterized iPSCs from a single individual [112,114], or were not able to distinguish between genetic and cell type of origin effects [115,116]. For example, though Kim et al. [116] reported a similar number of DM loci (137-370) between iPSCs derived from different cell types as we observed in our study, Kim et al. interpreted their observation as evidence for a marked effect of the donor cells. Yet, our observation that DNA methylation is quite homogenous across all iPSCs (both within replicates and between L-iPSCs and F-iPSCs; S8C/D Fig.), is not in disagreement with the observations of Kim et al.

Indeed, our study explicitly models the contribution of genetic background to variation in DNA methylation levels in iPSCs. When we consider DNA methylation in the context of variation explained by inter-individual differences,

we find a remarkably small effect associated with cell type of origin. Moreover, even unsupervised clustering (based on either DNA methylation or gene expression data) indicated that samples largely clustered by individual. We found little evidence of clustering by cell type of origin. When we turned our attention to individual loci, only 197 (0.043%) tested CpGs were classified as DM between L-iPSCs and F-iPSCs, compared with 190,356 (41.7%) loci that were classified as DM between LCLs and fibroblasts.

Our observation that only a handful of DM sites may drive regulatory differences between iPSCs from different origins is consistent with recent work by Rouhani and Kumasaka et al. [123] where a similar study design was employed examining only gene expression levels. Indeed, as in Kumasaka et al., we found that individual genetic background captures a much larger proportion of gene regulatory variation than cell type of origin using both the DNA methylation and gene expression data.

Future work needs to address additional pertinent questions. First, our study was limited to methylation and gene expression levels in iPSCs. Future studies should focus on additional epigenetic and regulatory markers. Second, we focused on regulatory differences between iPSCs, but did not study differentiated cell types. This needs to be addressed in the future because the degree to which iPSCs retain regulatory signatures of their cell type of origin ultimately is expected to influence the extent to which iPSCs can be used as a model system for studying complex traits in differentiated cell types.

In conclusion, our study demonstrated that when accounting for individual, the impact of cell type of origin on DNA methylation and gene expression in iPSCs is limited to a small number of CpGs, which cluster into an even smaller number of genomic loci, and a single gene, with almost no detectable influence genome-wide. Our observations further confirm the usefulness of iPSCs for genetic studies regardless of the original somatic cell type. The high correlation of DNA methylation and gene expression levels (S8C/D Fig.) between individuals, demonstrate the faithfulness of the model, though as we pointed out – similar studies in differentiate cells are required to generalize these conclusions. While cell type of origin should continue to be carefully documented, our data also suggest that future studies should focus on collecting more individuals rather than establishing multiple iPSC clones from the same individual.

3.5 Materials and Methods

Isolation and culture of fibroblasts and LCLs

Skin punch biopsies and blood were collected from the same individual within 20 minutes under University of Chicago IRB protocol 11-0524 (samples from four individuals were collected over three collection dates; samples from individuals 3 and 4 were collected on the same date). Skin and blood samples from an individual were processed at the same time. Fibroblast isolation and culture was conducted using the approach described in detail in Gallego Romero et al [127]. Briefly, skin punch biopsies (3mm) were digested using 0.5%

collagenase B (Roche), isolated fibroblasts were cultured in DMEM (Life Technologies) supplemented with 10% fetal bovine serum (FBS; JR Scientific), 0.1mM NEAA, 2mM GlutaMAX (both from Life Technologies), 1% penicillin/streptomycin (Fisher), 64mg/L L-ascorbic acid 2-phosphate sesquimagnesium salt hydrate (Santa Cruz Biotechnology), at 5% CO_2 and 5% O_2.

All other cell culture was performed at 5% CO_2 and atmospheric O_2. For LCL generation, whole blood was drawn (within 20 minutes of obtaining skin punch biopsies) into two 8.5mL glass yellow top tubes (Acid Citrate Dextrose Solution A tubes; BD). Blood tubes were stored at room temperature and processed within 12 hours of collection. To isolate lymphocytes, we diluted whole blood with an equal amount of RPMI 1640 (Corning), diluted blood was slowly layered onto Ficoll-Paque (GE Lifescience) in 50 mL centrifuge tubes. This gradient was centrifuged at 1700 rpm for 30 minutes without acceleration or braking. Leukocytes and platelets formed a white band at the interface between the blood plasma and the Ficoll (called the buffy coat). We collected the buffy coat using a Pastette® and to that added 10mL of PBS. The collected buffy coat was then washed three times with PBS.

For EBV transformation, 4 x 10^6 fresh lymphocytes collected as described above were resuspended in a total of 4.5 ml of RPMI 1640 culture medium (Corning) containing 20% FBS and 1:100 phytohemagglutinin (PHA-M; LifeTechnologies) and transferred to a T-25 flask. EBV supernatant produced by the B95-8 cell lines (provided by the Ober lab) was added at 1:10 to the culture

flask. Cells were left undisturbed for three to five days before adding fresh media. Flasks were subsequently examined weekly for changes in cell growth as indicated by acidic pH (yellow color) and the appearance of clumps of cells growing in suspension. Once growth was established (21-35 days), cells were diluted or split to several flasks. When the cell density reached 8 x 10^5 to 1 x 10^6 cells per mL they were cryopreserved at a density of 10 x 10^6 cells per ml of freezing media in cryovials. All LCLs using this study were transformed with the same lot of EBV supernatant.

Episomally-reprogrammed iPSCs

To establish iPSCs we transfected LCLs (Amaxa™ Nucleofector™ Technology; Lonza) and fibroblasts (Neon® Transfection System; Life Technologies) with oriP/EBNA1 PCXLE based episomal plasmids that containing the genes *OCT3/4, SOX2, KLF4, L-MYC, LIN28*, and an shRNA against *p53* [121]. We supplemented these plasmids with an *in vitro*-transcribed EBNA1 mRNA transcript to promote exogenous vector retention following electroporation of the episomal vector [128,129]. Fibroblasts from all individuals were reprogrammed in two batches. LCLs were reprogrammed in four batches. The first three batches contained LCLs from all four individuals. Individual 4 failed reprogramming in batches one and three. A final fourth batch was therefore done with only individual 4. We plated a range of 10,000 - 40,000 transfected cells per well in a 6-well plate. Within 21 days colonies were visible and manually passaged onto a fresh plate of irradiated CF1 mouse embryonic fibroblasts

(MEF). We passaged these new iPSC colonies on MEF in hESC media (DMEM/F12 (Corning) supplemented with 20% KOSR (LifeTechnologies), 0.1mM NEAA, 2mM GlutaMAX, 1% Pen/Strep, 0.1% 2-Mercaptoethanol (LifeTechnologies)). Fibroblast derived iPSCs were supplemented with 100ng/mL human basic fibroblast growth factor, versus 25ng/mL for LCL derived iPSCs; all other culture conditions were identical. After 10 passages of growth we transitioned the cultures to feeder-free conditions and cultured them for an additional three passages before collecting cell pellets for analysis. Feeder-free cultures were grown using 0.01mg/cm^2 (1:100) hESC-grade Matrigel (BD Sciences) and Essential 8 media (LifeTechnologies). Passaging was done using DPBS supplemented with 0.5mM EDTA. All RNA and DNA were isolated using Zymo dual extraction kits (Zymo Research) with a DNase treatment during RNA extraction (Qiagen).

Characterization of iPSCs

All iPSC lines were characterized as described previously [127]. Briefly, we initially confirmed pluripotency using PluriTest [130], a classifier that assigns samples a pluripotency score and novelty score based on genome-wide gene expression data. All samples were classified as pluripotent and had a low novelty score (S1 Fig.). We next performed qPCR using 1 μg of total RNA, converted to cDNA, from all samples to confirm the endogenous expression of pluripotency genes: *OCT3/4*, *NANOG*, and *SOX2* (S2A-C Fig.). Additionally, we tested for the presence and expression of the EBV gene *EBNA-1* using PCR (S2D/3 Figs.). We

tested all samples for both genomic integrations and vector-based EBV. We did this using primers designed to amplify the *EBNA-1* segment found in both the episomal vectors and the EBV used to transform LCLs. If the cell was positive (a single positive case was found: Ind4 F-iPSC), we further tested the origin of the EBV (genomic or episomal) using primers specific to the *LMP-2A* gene found in EBV or part of the sequence specific to the episomal plasmid (S3 Fig.). Finally, we confirmed the ability of all iPSC lines to differentiate into the three main germ layers using the embryoid body (EB) assay. The EBs were imaged for the presence of all three germ layers (S4 Fig.). It should also be noted that gene expression and DNA methylation levels are extremely similar between iPSC lines. This relative homogeneity further demonstrates the quality of our iPSC lines. In summary, all iPSC lines established in this study showed expression of pluripotent genes quantified by qPCR, generated EBs for all three germ layers, and were classified as pluripotent based on PluriTest.

Processing of methylation array

Extracted DNA was bisulphite-converted and hybridized to the Infinium HumanMethylation450 BeadChip (Illumina) at the University of Chicago Functional Genomics facility. To validate the array probe specificity, probe sequences were mapped to an *in silico* bisulfite-converted genome using the Bismark aligner [131]. Only probes that mapped uniquely to the human genome were retained (n = 459,221). We further removed data from probes associated with low signal (detection *P-value > 0.01*) in more than 25% of samples (retained

data from n = 455,910 loci). Raw output from the array (IDAT files) were processed using the minfi package [124] in R.

We performed standard background correction as suggested by Illumina [124], and corrected for the different distribution of the two probe types on the array using SWAN [125] (S5 Fig.). Additionally, we quantile normalized the red and green color channels (corresponding to methylated and unmethylated signal respectively) separately (S6A/B Fig.). To calculate methylation levels (reported as β-values) we divided the methylated signal by the total signal from both channels. β-values were considered estimates of the fraction of alleles methylated at that particular locus in the entire cell population.

Processing of expression arrays

RNA quality was confirmed by quantifying sample's RNA Integrity Number (RIN) on an Agilent 2100 Bioanalyzer (Agilent Technologies). All samples had a RIN of 10. The extracted RNA from all samples was hybridized to the Illumina HT12v4 Expression BeadChip array (Illumina) at the University of Chicago Functional Genomics facility. Sample processing was performed using the lumi package in R [132]. We excluded data from a subset of probes prior to our analysis: First, we mapped the probe sequences to the human genome hg19 and kept only those with a quality score of 37, indicative of unambiguous mapping (n = 40,198; note that we also explicitly pre-filtered the 5,587 probes which were annotated as spanning exon-exon junctions to avoid mapping errors). Second, we downloaded the HapMap CEU SNPs

(http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/2010-08_phaseII+III/forward/) and converted their coordinates from hg18 to hg19 using the UCSC liftOver utility [133]. We retained only those probes that did not overlap any SNP with a minor allele frequency greater than 5% (n = 34,508). Third, we converted the Illumina probe IDs to Ensembl gene IDs using the R/Bioconductor package biomaRt [134] and retained only those probes that are associated with exactly one Ensembl gene ID (Ensembl 75 - Feb 2014; n = 22,032). The full pipeline was implemented using the Python package Snakemake [135]. We defined a gene as expressed in a given sample if at least one probe mapping to it had a detection *P-value* < 0.05. In the case of L-iPSCs, we defined a gene as expressed in an individual if any associated probes had a detection *P-value* < 0.05 in at least one biological replicate. Using these criteria, we identified all genes expressed in at least three individuals in at least one cell type (S7 Fig.; n = 14,111 probes associated with 11,054 annotated genes). In the case that multiple expressed probes were associated with the same ENSEMBL gene (n = 3,057), we only retained data from the 3'-most detected probe. Following these filtration steps, we obtained estimates of expression levels in all samples across 11,054 genes. Data from the 11,054 genes were quantile normalized using the lumiExpresso function in lumi [132] (S6C/D Fig.).

Unsupervised hierarchical clustering and heatmaps

Only data from autosomal probes were retained for the hierarchical clustering analyses in order to reduce bias towards clustering by individual or sex

(n = 10,648 expression, and n = 445,277 methylation). We calculated a matrix of pairwise Euclidean distances between samples from the methylation and expression data separately. From these matrices we performed hierarchical clustering analyzing using the complete linkage method as implemented in the R function hclust. The observed dendrograms remained consistent regardless of the linkage method chosen (complete, single, or average). The 1,000 most variable loci were defined by taking the loci with the highest variance in iPSCs. Clustering based on the 1,000 most variable probes were processed in an identical manner as above. Heatmaps were generated from matrices of pairwise Pearson correlations between samples using data from autosomes and sex chromosomes.

Analysis of differences in gene expression and methylation levels

Data from probes on both autosomes and sex chromosomes were included in this analysis, given that individuals were balanced across cell types (n = 455,910 CpGs; n = 11,054 genes). Additionally, we anticipated that sites on the sex chromosomes may be particularly sensitive to mis-regulation during reprogramming [136]. Differential expression and methylation analyses were performed using linear modeling and empirical Bayes methods as implemented in the limma package [126]. We tested for differential methylation and expression, using locus-specific models, between L-iPSCs and F-iPSCs; L-iPSCs and LCLs; F-iPSCs and fibroblasts; and between fibroblasts and LCLs. We considered a locus DM or DE at an FDR < 5% (Benjamini Hochberg). We

also tested for DE genes between L-iPSCs and F-iPSCs using only genes that were classified as DE between L-iPSCs and LCLs; F-iPSCs and fibroblasts; and LCLs and fibroblasts (S11 Fig.). We estimated FDRs separately each time we considered only subsets of the data.

Due to the imbalance of L-iPSC samples to F-iPSC samples we repeated our analyses using data from a reduced set of samples. Namely, we randomly sampled a single replicate of the L-iPSC from each individual. As expected, reducing the number of L-iPSC samples greatly reduces the number of loci classified as DM between L-iPSCs and F-iPSCs as well as between L-iPSCs and LCLs. However, the number of DM loci was reduced across all other contrasts as limma models the entire matrix together (S12 Fig.). Interestingly, we found that different combinations of replicates yielded DE genes other than *TSTD1*. Therefore, we sampled all possible combinations and overall, found six genes that were classified as DE (FDR 5%) in at least one of the combinations of reduced samples. Of note, we never classify *TSTD1* as DE (FDR 5%) in the reduced data set. The most common DE gene, *INPP5F*, is the only gene that also has nearby DM CpGs (five of the 25 nearby loci). Additionally, in the full model, *INPP5F* has the second lowest *P value* (uncorrected $P = 6.84 \times 10^{-5}$; FDR 38%). However, *INPP5F* was not DE between LCLs and fibroblasts, but was DE between LCLs and L-iPSCs and also fibroblasts and F-iPSCs (S3A-D Tables; Fig. 3C).

Enrichment of DM loci in regulatory and genomic features

We employed two strategies to identify enrichments of DM loci between L-iPSCs and F-iPSCs in regulatory features. First, we used the regulatory states defined by Ernst et al. [137]. We tested for enrichments in all regulatory categories using a χ-square test comparing the number DM loci and total probes within each regulatory class to the number DM loci and total probes outside the regulatory class. We found no significant enrichment for any of the defined regulatory states.

Next, we used the UCSC_RefGene_Group annotation as supplied by Illumina. These annotations detail the location of probes in relation to genes (1st Exon, 3' UTR, 5' UTR, Gene Body, within 1.5kb of a TSS or within 200bp of a TSS). We identified significant enrichments of DM loci within 1.5kb of a TSS and gene bodies. However, there are six probes classified as both within a gene body and within 1.5kb of a TSS. We chose to report both results because it is difficult to deconvolute these categories.

We also considered the position of DM loci in relation to genes. The annotations were defined by Illumina. We were able to identify 37 genes associated with DM loci, but we only had corresponding gene expression data for 24 of these genes. We attempted to identify signals of enrichment in DE levels between L-iPSCs and F-iPSCs in these 24 genes. To this end, we compared the log fold changes in gene expression between L-iPSCs and F-iPSCs from genes with nearby DM loci between L-iPSCs and F-iPSCs to 10,000 random samplings of log fold change in expression between L-iPSCs and F-iPSCs from all genes and found no enrichment for increased log fold changes.

Proportion of variance explained

To estimate the proportion of variance explained by individual and cell type of origin we performed a linear mixed model with a fixed effect for cell type of origin and a random effect for individual. Only data from autosomes were included in this analysis so that the results would not be biased toward differences in individuals (n = 10,648 expression, and n = 445,277 methylation). To calculate the proportion of variance explained we divided the variance components of each term by the total variance in gene expression (Fig. 4). When focusing on CpGs and genes with previously identified genetic associations (eQTLs and meQTLs, respectively) we used genes with at least one eQTL identified by Lappalainen et al. 2013 [138] and CpGs with at least one meQTL identified by Banovich et al. 2014 [36] (S11 Fig.).

Accession numbers

The expression and methylation data sets supporting the results of this article are available in the Gene Expression Omnibus (GEO) under accession GSE65079 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65079).

Ethics, consent and permissions

All individuals consented to study participation under University of Chicago IRB protocol 11-0524.

3.6 Appendix B: Supplementary Materials

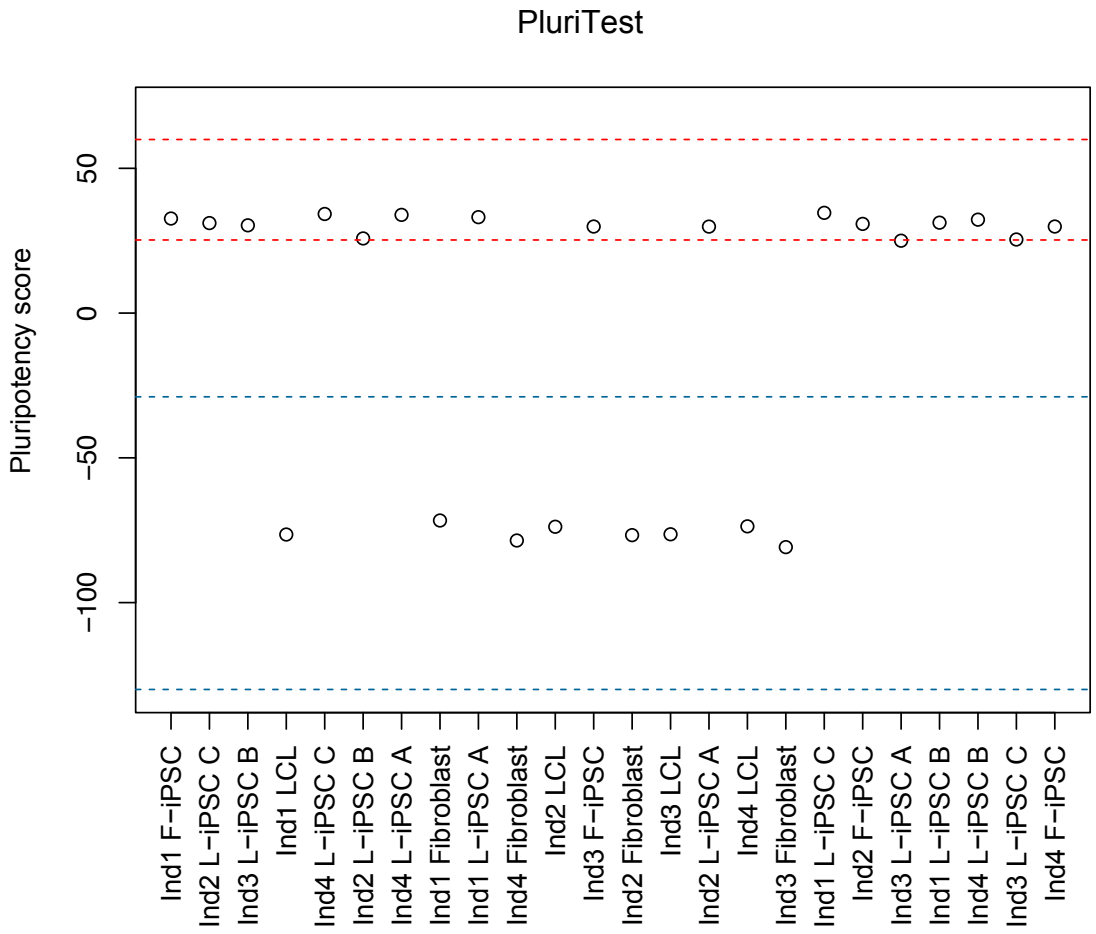

Figure S3.1 Quality control of iPSCs.
iPSC lines QC - PluriTest pluriscore results for all samples, showing all iPSC samples fall within the pluripotent threshold (red dashed lines). Additionally, all primary tissue samples fall within the non-iPSC cell type classification (blue dashed lines).

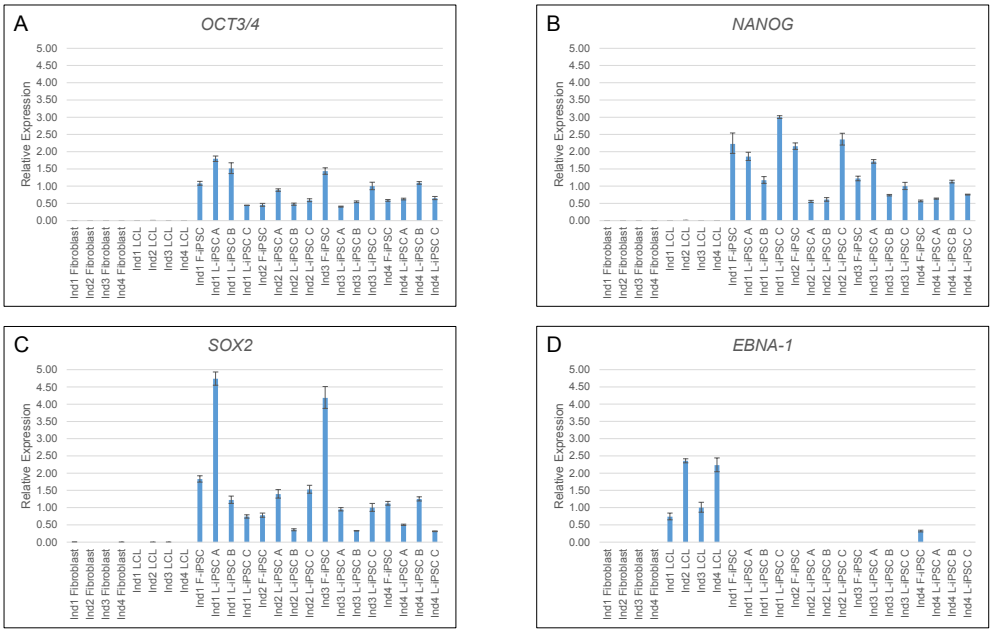

Figure S3.2 Quality control of iPSC Lines.
iPSC lines QC - Quantitative PCR (qPCR) of pluripotency genes (a) *OCT3/4*, (b) *NANOG*, and (c) *SOX2* normalized on randomly selected Ind3 L-iPSC C. Relative expression is the RQ value with respect to *GAPDH* expression, with error bars representing the calculated min and max RQ value. All iPSC lines show endogenous expression of these pluripotency genes. (d) Expression of *EBNA-1*, a required viral gene of Epstein-Barr virus (EBV), normalized on randomly selected Ind3 LCL. *EBNA-1* expression could stem from either the reprogramming vectors or, in LCLs and L-iPSCs, expression of integrated genomic EBV. Ind4 F-iPSC shows low expression of EBNA-1 due to low retention of reprogramming vectors as confirmed in Supplementary Fig. 3. This sample is kept for data analysis because all other QC measures are met and the sample is not an outlier in overall gene expression or DNA methylation.

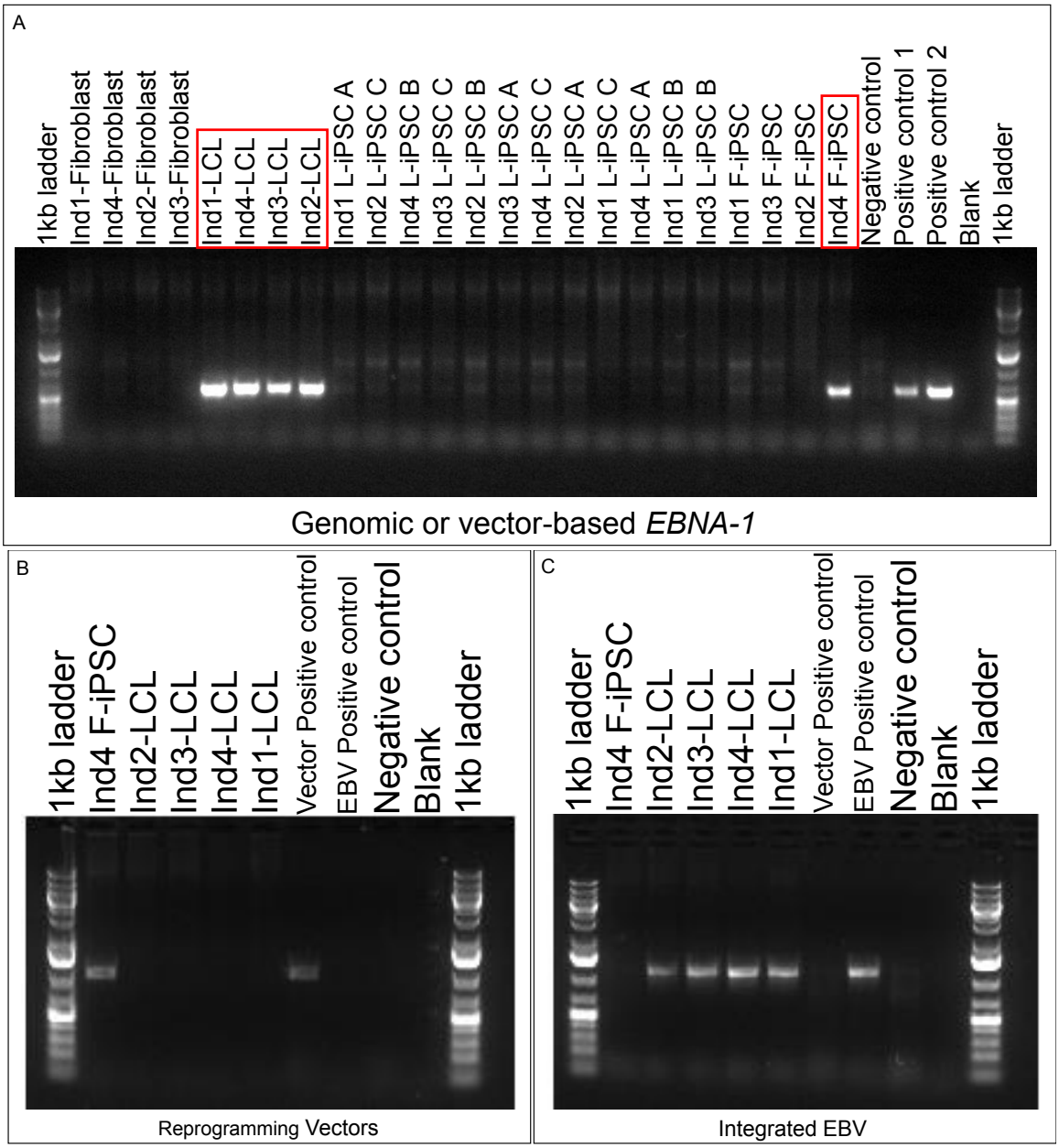

Figure S3.3 Quality control of iPSCs.
(a) PCR on DNA for presence or absence of EBV, both integrated and non-integrated (reprogramming vector based). All four LCLs showed the presence of EBV along with one iPSC line, Ind4 F-iPSC. Additional banding in the images is due to RNA in the sample. These five samples, highlighted by a red box, were taken forward for two additional PCRs. First, the five samples were tested for the presence of the reprogramming vectors (b), of which only Ind4 F-iPSC was positive. Lastly, the five samples were tested for EBV based on the presence of the *LMP-2A* sequence (c; an EBV gene not found on the reprogramming vector). All LCLs were positive for EBV, and the iPSC sample was not.

Figure S3.4 Quality control of iPSCs.

iPSC lines QC - Embryoid body (EB) formation from iPSC lines to validate the ability to differentiate into all three germ layers. The leftmost column (a) shows EBs stained with Nestin, a cytoplasmic stain for ectoderm in green and MAP2, a cytoplasmic stain for ectoderm in red. The center column (b) shows EBs stained with SMA, a cytoplasmic stain for mesoderm in green and again for MAP2 in red. The rightmost column (c) shows EBs stained with AFP, a cytoplasmic stain for endoderm in green and HNF3β, a nuclear stain for endoderm in red. All iPSC lines generated showed the ability to differentiate into all three germ layers. All

imaging was done at 10x magnification and nuclei were stained blue with Hoechst.

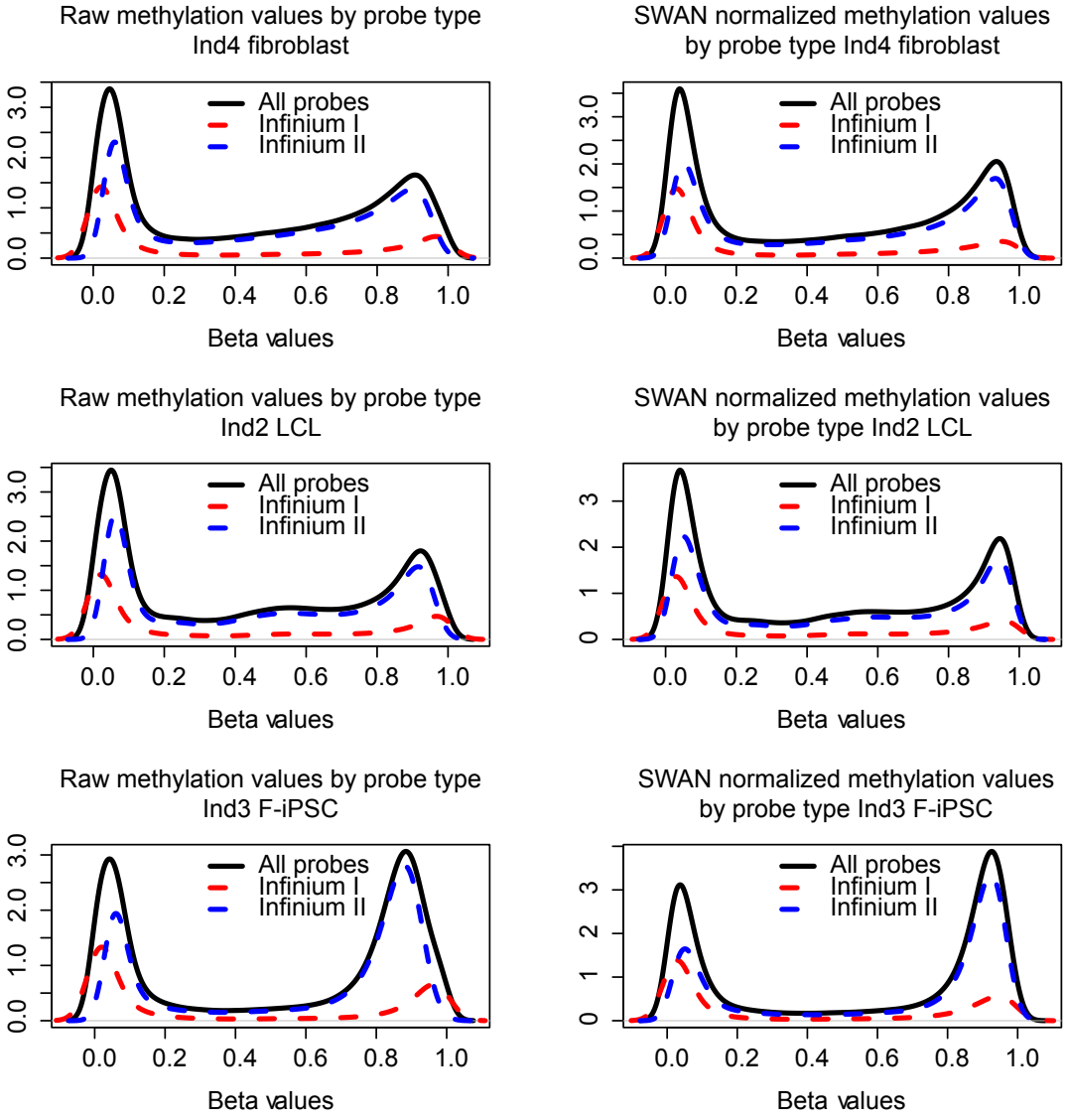

Figure S3.5 DNA methylation density plots.
Representative density plots of DNA methylation levels separated by type I and type II probes before and after SWAN Normalization.

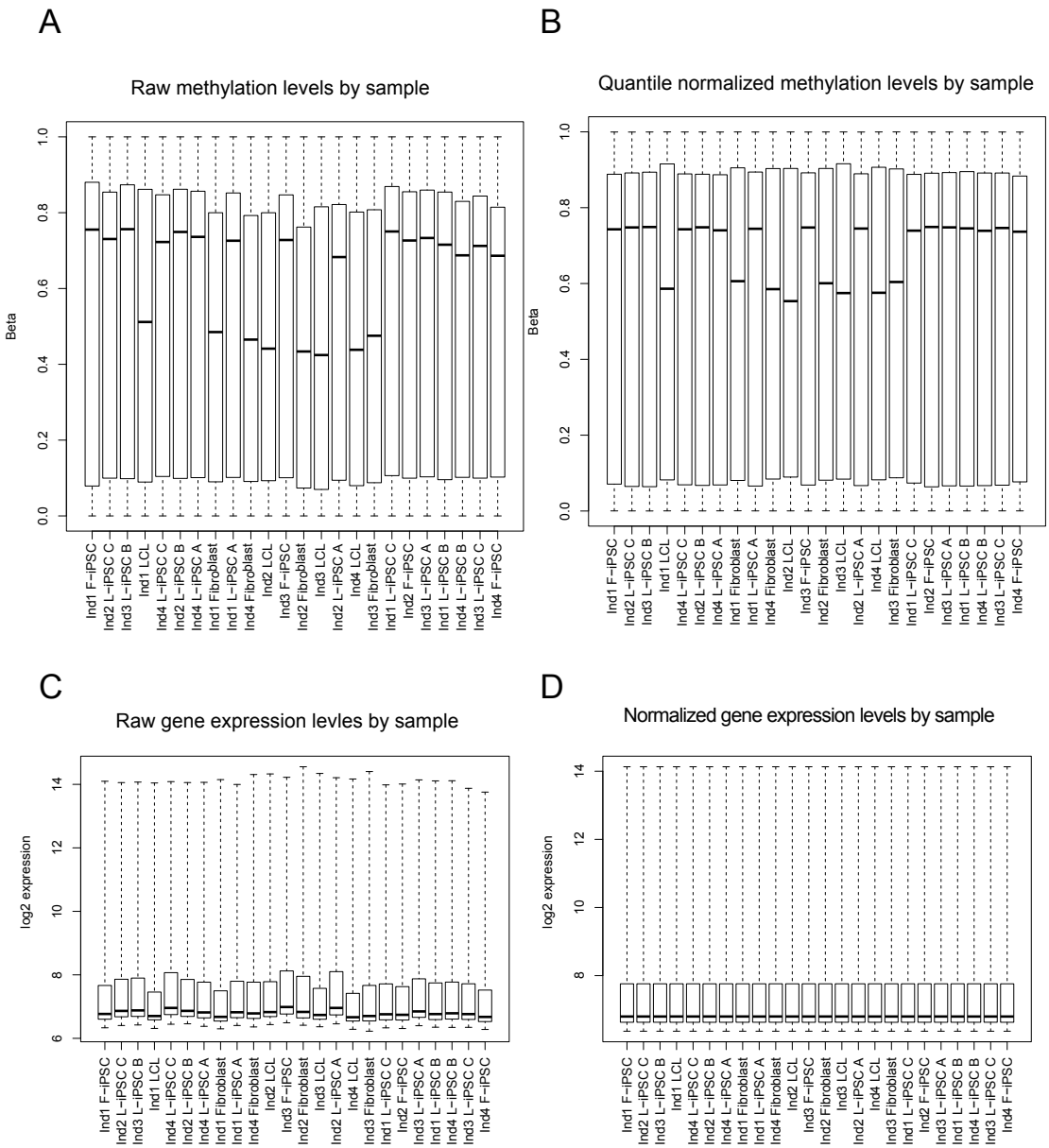

Figure S3.6 Array data normalization.
Methylation levels (Beta) (a) pre- and (b) post- quantile normalization. Quantile normalization was performed independently on the red and green color channels. Gene expression data (c) pre- and (d) post- quantile normalization.

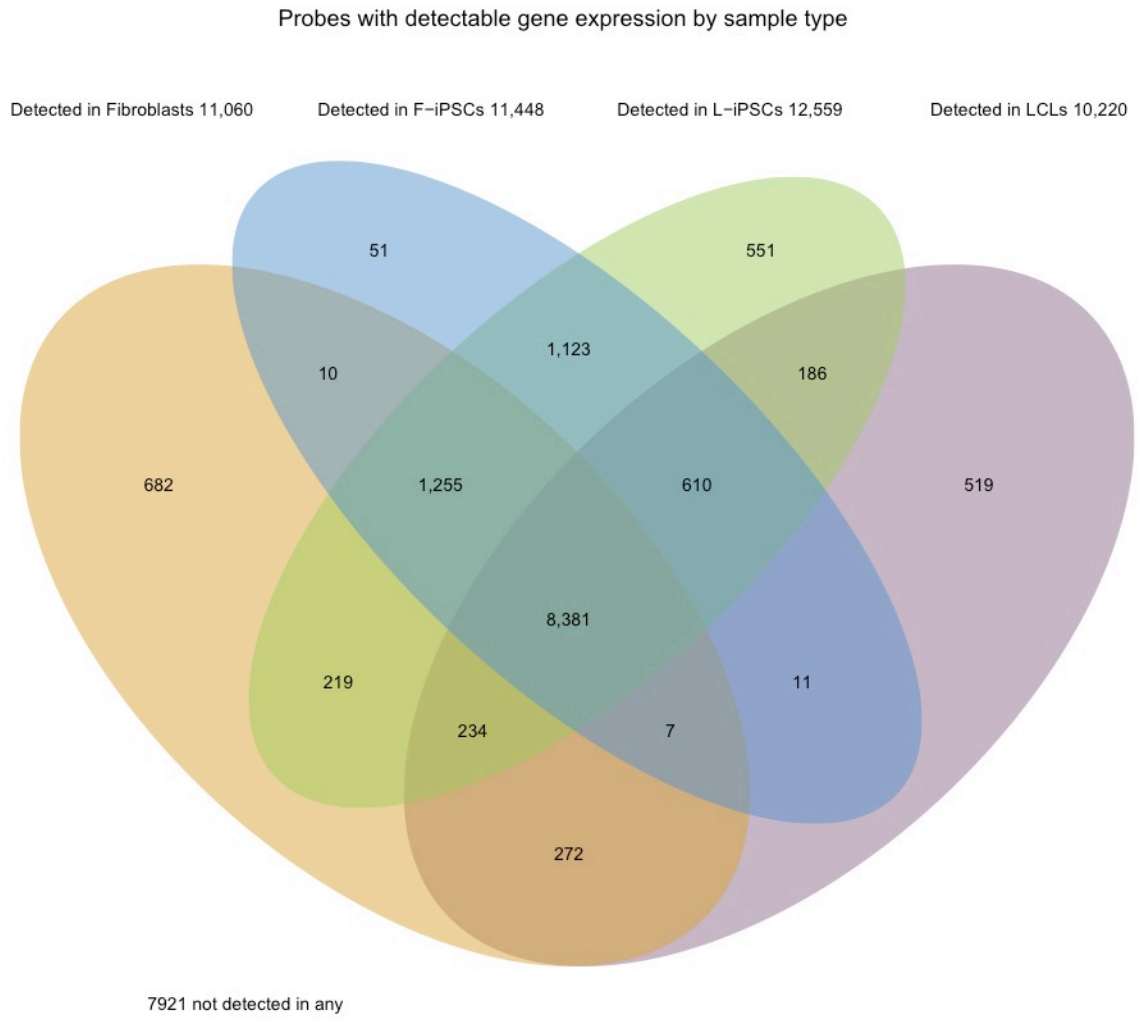

Figure S3.7 Probe inclusion scheme.
For 22,032 probes we defined a gene as expressed in a given sample if at least one probe mapping to it had a detection P-Value < 0.05. In the case of L-iPSCs, we defined a gene as expressed in an individual if any associated probes had a detection P-Value < 0.05 in at least one biological replicate. Using these criteria, we identified all genes expressed in at least three individuals in at least one cell type (n = 14,111 probes, associated with 11,054 genes).

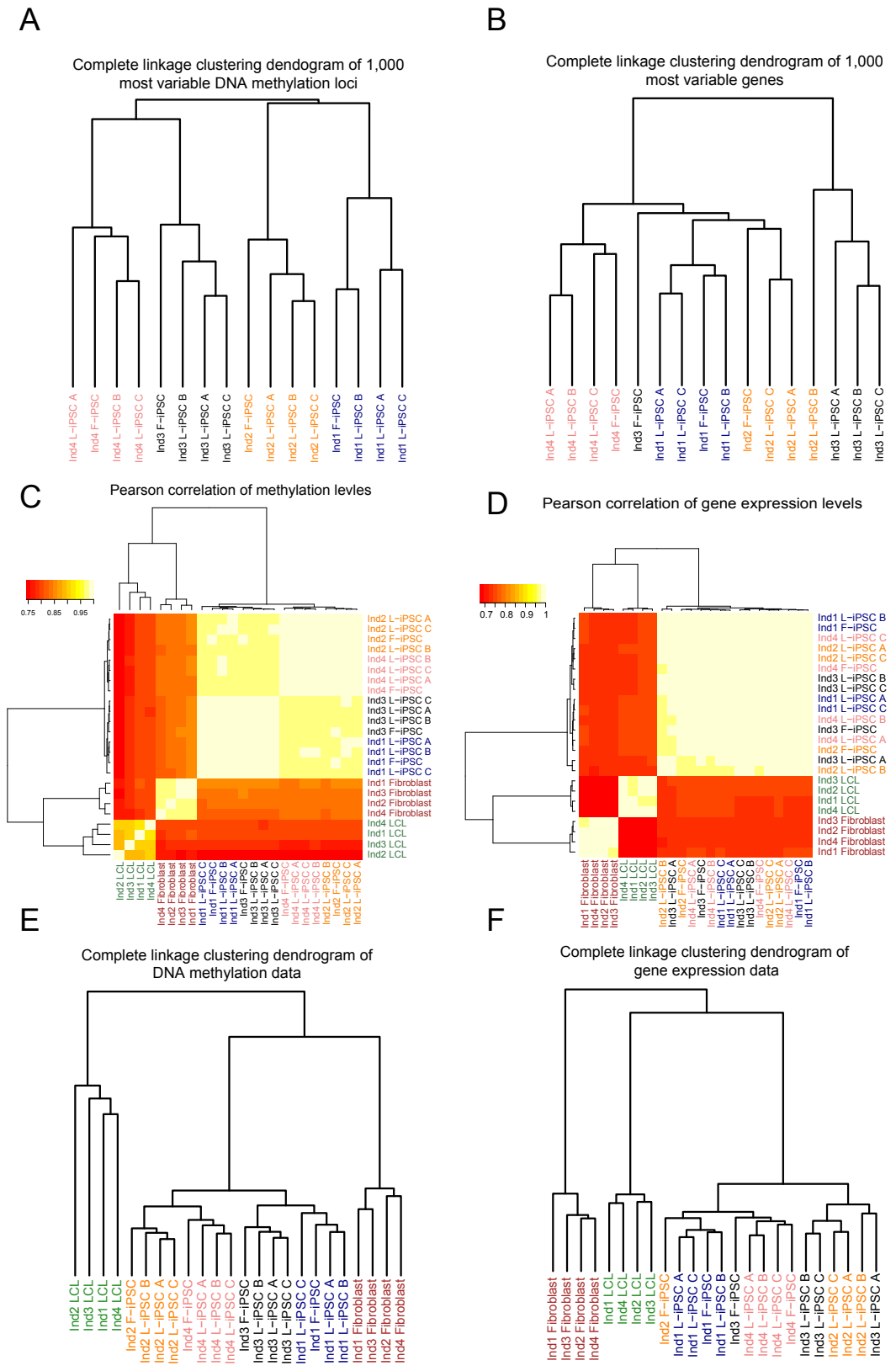

Figure S3.8 Hierarchical clustering.

Hierarchical clustering using the complete linkage method and Euclidean distance from the 1,000 most variable autosomal iPSC loci for (a) methylation data and (b) expression data. Heatmap showing pairwise Pearson correlations between all samples for all loci (autosomes and sex chromosomes) (c) methylation data and (d) gene expression data: note all iPSCs are highly correlated. Hierarchical clustering using the complete linkage method and Euclidean distance from all loci (autosomes and sex chromosomes) for (e) methylation data (n = 455,910) and (f) gene expression data (n = 11,054).

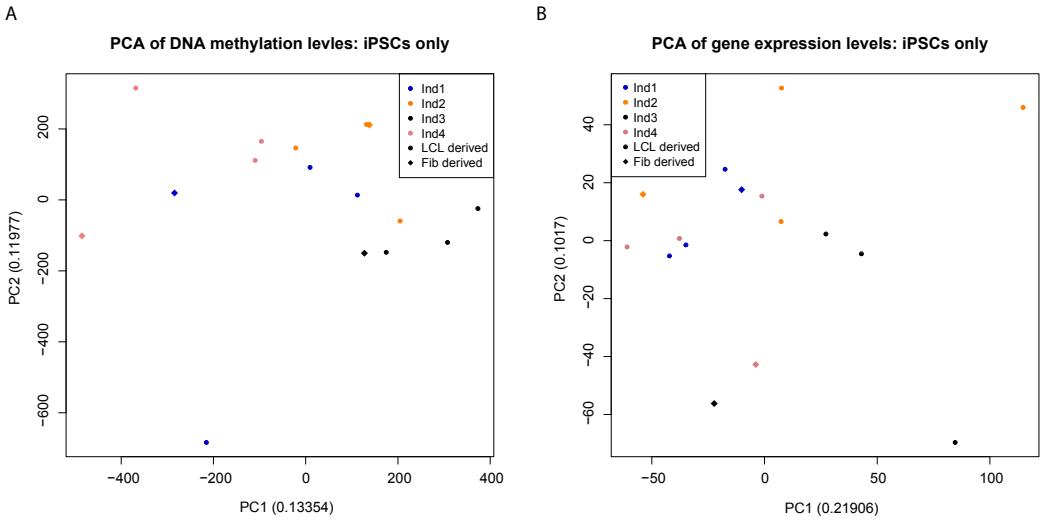

Figure S3.9 Principal components analysis (PCA).
Results of PCA on (a) methylation levels and (b) gene expression levels, using only autosomal loci in the iPSC samples.

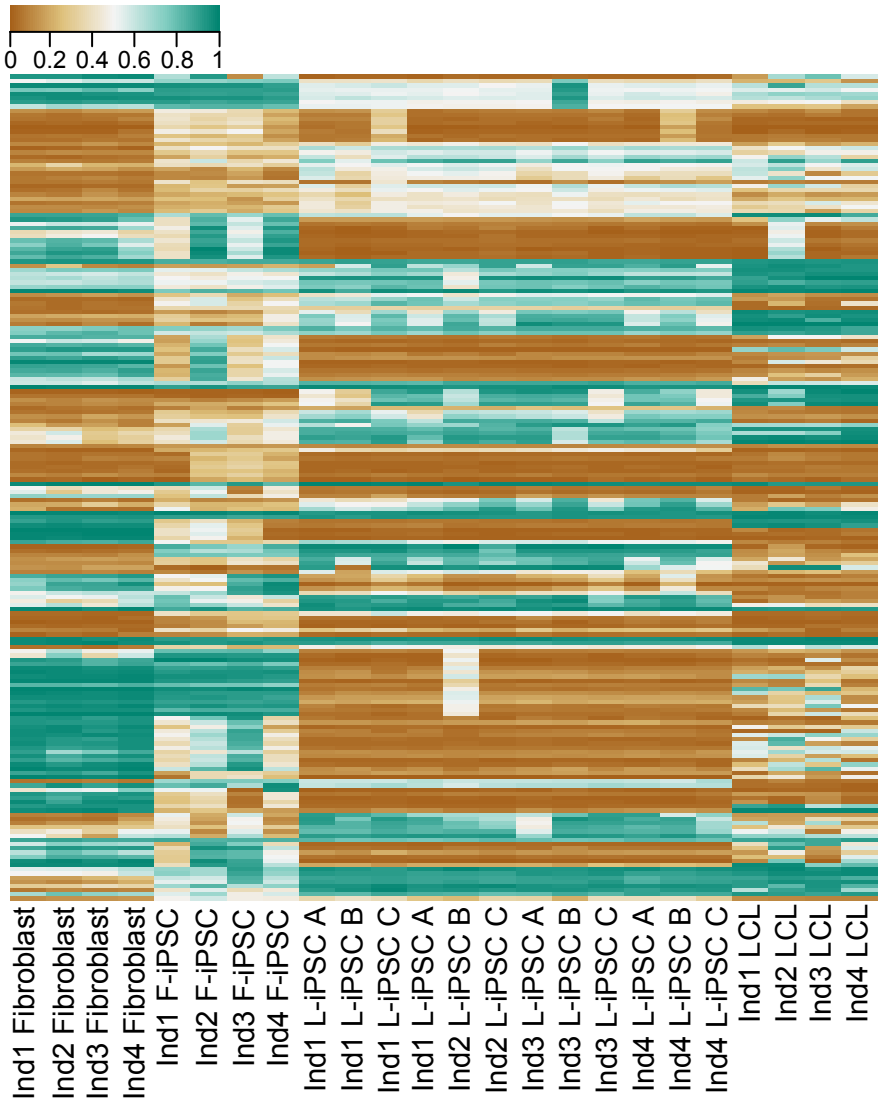

Figure S3.10 Heatmap of DM loci.

A heatmap of methylation levels at loci DM between L-iPSC and F-iPSC (n = 197), ordered by genomic location.

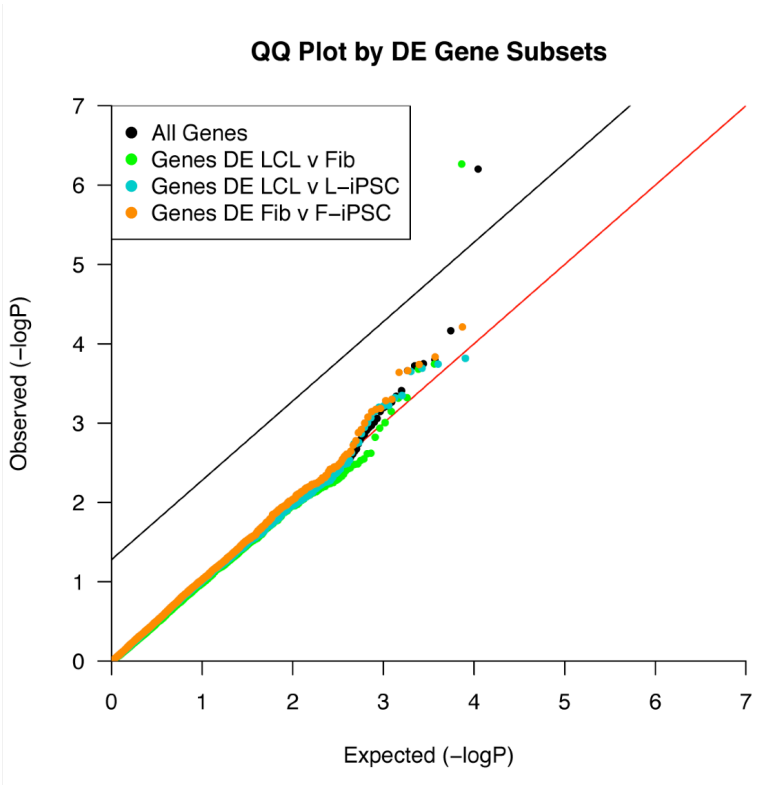

Figure S3.11 DE tests in gene subsets.
To confirm that the test to detect DE genes was not underpowered, we also tested for DE in subsets of genes most likely to be DE between L-iPSC and F-iPSC – genes that were identified as DE in the other contrasts tested. We found no enrichment of significant P-Values based on DE tests with these subsets; see QQ plot of P-Values considering DE tests between L-iPSCs and F-iPSCs using four distinct gene sets: all genes, only genes DE between LCL and fibroblasts, only genes DE between LCL and L-iPSCs, and only genes DE between fibroblasts and F-iPSCs.

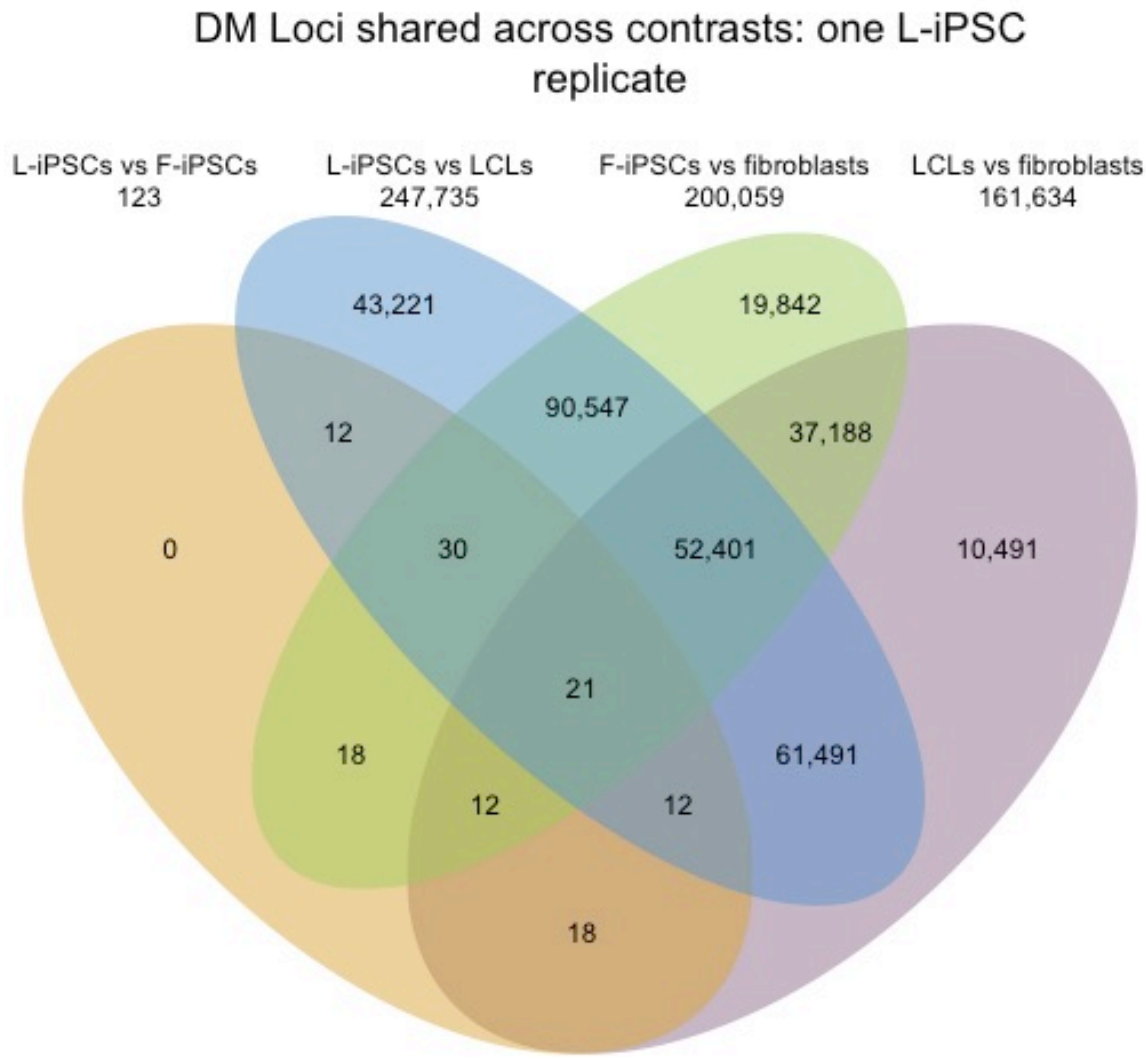

Figure S3.12 Differential methylation with single L-iPSC replicate.
A Venn diagram depicting differentially methylated (DM) loci identified at an FDR of 5% overlapping between different contrasts with only a single L-iPSC replicate from each individual. A general decrease in the number of DM loci is observed across all contrasts as limma models all the data together. Yet, a far more marked decrease in the number of DM loci is observed in contrasts containing L-iPSCs.

CHAPTER 4: HUMAN INDUCED PLURIPOTENT STEM CELLS: A POWERFUL MODEL TO INVESTIGATE INTER-INDIVIDUAL REGULATORY VARIATION ACROSS CELL TYPES

4.1 Abstract

Human induced pluripotent stem cells (iPSCs) provide a powerful system to study complex human traits. To investigate inter-individual variation in gene regulation across multiple cell types from the same individuals, we established and validated a panel of 59 iPSCs from lymphoblastoid cell lines (LCLs) of Yoruba individuals, which have been extensively studied in the past. The genome sequences of all individuals were also available to us. We collected RNA sequencing, chromatin accessibility, and DNA methylation data from the LCLs and the iPSCs, as well as RNA sequencing from iPSC-derived cardiomyocytes (iPSC-CMs) from 13 of the same individuals.

Using these gene regulatory data, we identified thousands of genetic associations with inter-individual variation in gene expression levels (eQTLs), methylation levels (meQTLs), and chromatin accessibility (caQTLs), across cell types. We found that regulatory variation is lower in iPSCs compared with the differentiated cell types, consistent with the intuition that developmental processes are generally canalized. By considering transcription factor footprints and inferred chromatin states, we were able to provide putative mechanistic explanations for many differences in regulatory QTL associations across cell types. In particular, we identified a large number of cell type specific regulatory QTLs in distal enhancers, which are likely to regulate tissue-specific gene expression patterns.

This study demonstrates the power of the iPS cellular model to dynamically study inter-individual variation in gene regulation.

4.2 Introduction

Understanding the genetic underpinnings of complex traits remains one of the major goals in human genetics. The advent of high throughput genotyping technologies (array and sequencing based) represented a transformative period in the study of complex traits. Researchers postulated that with large samples of individuals and well executed case control studies we would identify the majority of genetic drivers of complex traits including disease [3]. Unfortunately, it became clear that complex traits were even more complex then originally believed. Recent large-scale meta genome-wide association studies (GWAS) with traits such as BMI [12] suggest that there may be thousands of genetic variants with small effect sizes contributing to complex traits. Within the current framework prohibitively large sample sizes would be needed to fully elucidate the genetic architecture of any complex trait. However, GWAS studies have provided a wealth of information about the general properties of loci affecting complex traits. Notably, the majority of such loci lie outside of genes and likely act by modifying gene expression [17]. Indeed, recent work has shown you can dramatically increase your ability to identify genetic variants associated with disease traits by incorporating gene expression data from a disease relevant tissue [42]. These

results demonstrate the importance of studying gene regulation in identifying genetic variants associated with complex traits.

To this end many studies have examined the effect of genetic variation on gene expression [31-33] and other regulatory phenotypes [27,34-41]. However, due to ethical and practical constraints, these studies have been limited to commercially available cell lines [27,34-37], easily accessible tissues (eg. skin and blood), and, more recently, post-mortem tissues [32]. While these studies have provided valuable insight into the genetic architecture of gene regulation, none of the aforementioned models provide a flexible framework to study inter-individual variation in gene regulation in multiple cell types from the same individual.

The discovery that somatic cells could be transformed into embryonic-like cells [43-45] and then re-differentiated into somatic cell types from any germ layer [46] provides a powerful cellular model to study gene regulation. Importantly, induced pluripotent stem cells (iPSCs) can be efficiently generated with a small number of exogenous factors [121]. Moreover, while the equivalence of iPSCs and embryonic stem cells (ESCs) remains debated, recent work using well-matched lines suggests that iPSCs are nearly indistinguishable from ESCs [139]. Recent work examining gene regulation in iPSCs has demonstrated that variation in gene expression and DNA methylation [123,140] is highly dependent on donor individuals. These results suggest that iPSCs can be used to study genetic effects on gene regulation. Indeed, one study has established that common genetic variation is associated with changes in gene expression and the

main driver of expression differences in iPSCs [141]. However, a more extensive evaluation of variation in multiple regulatory phenotypes from iPSCs and iPSC-derived cell types is lacking.

To this end we have generated a panel of iPSCs from 59 well characterized immortalized lymphoblastoid cell lines (LCLs). We have collected gene expression, chromatin accessibility, and DNA methylation data from this panel. Additionally, we have differentiated 13 of these lines into iPSC-derived cardiomyocytes (iPSC-CMs) from which we have collected gene expression. This study is the deepest characterization of gene regulation in iPSCs to date and represents a large advance in our ability to study the genetic architecture of gene regulation across cell types.

4.3 Results

Generation of high quality iPSCs from 59 Yoruba individuals

We successfully generated iPSCs from 59 Yoruba individuals (see methods). Briefly, LCLs were reprogrammed using a previously described episomal approach [121]. After a week in suspension culture cells were seeded onto a layer of gelatin and mouse embryonic fibroblasts. A single clonal colony is obtained from each line and passaged for ten weeks before final characterization and collection. Pluripotency and stability were confirmed using three methods. First, iPSCs were allowed to form embroyid bodies and then spontaneously differentiate. After a week of differentiation iPSCs were stained for tissues from

all three germ layers (Fig. 4.1A). Next, we applied a bioinoformatic classifier,

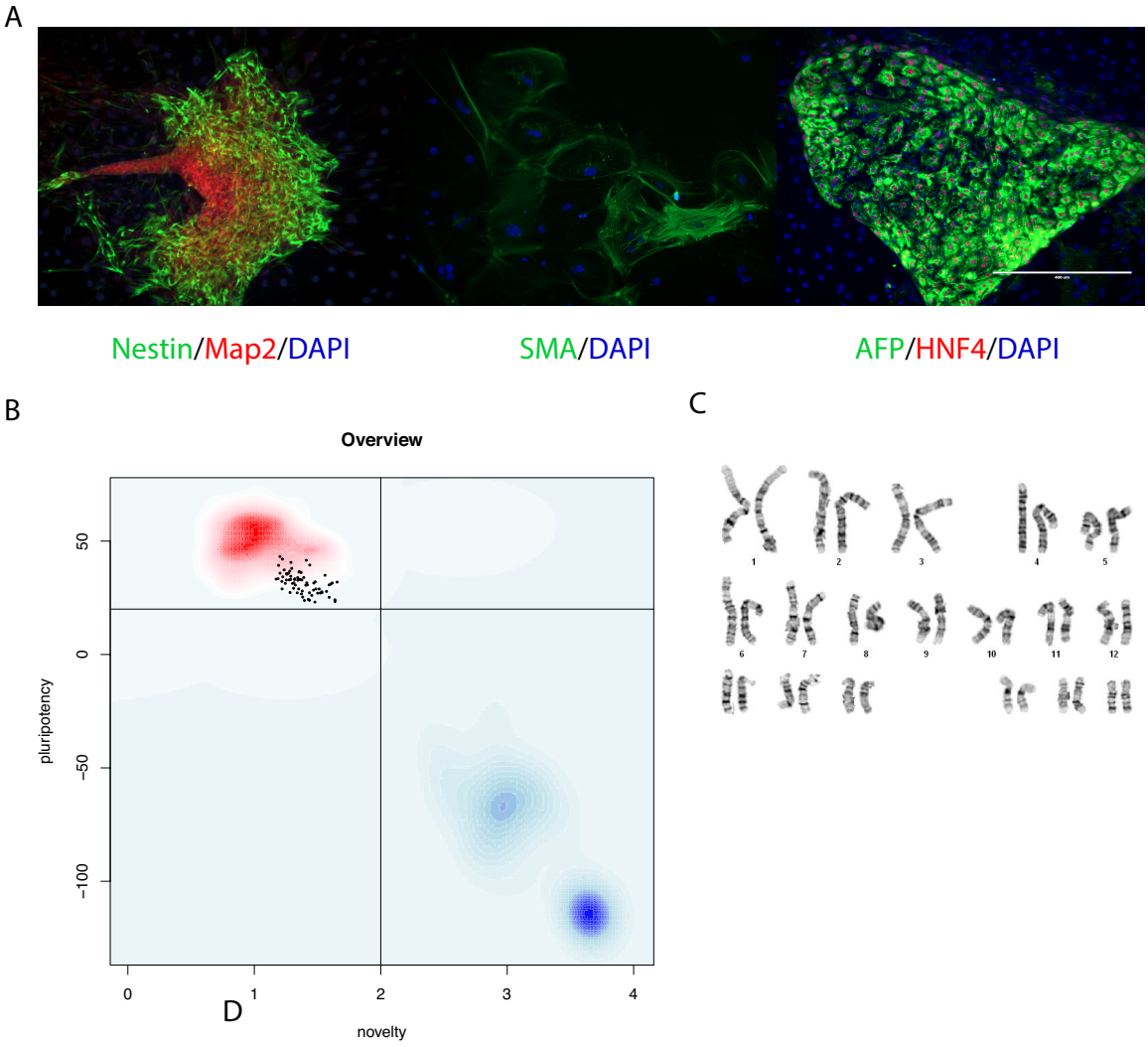

Figure 4.1 Quality control of iPSC lines

A) A representative image of immunohistochemistry staining for ectoderm, mesoderm, and endoderm cell types. B) Pluritest results. Upper left quadrant represents empirical cutoffs. C) A representative karyotype result

PluriTest [130], to our data. The classifier compares gene expression levels from uncharacterized lines to a “gold standard” panel of embryonic stem cells and iPSCs. Two metrics are obtained from this method providing information about the similarity in gene expression of canonical pluripotency genes and amount of aberrant unexpected expression (Fig. 4.1B). Finally, all lines were karyotyped to

demonstrate genomic stability (Fig. 4.1C). The iPSCs described here passed all quality controls and have been grown for at least 20 passages in a feeder culture system. Additionally, all lines are able to transition to feeder-free conditions using a commercially available growth medium and extracellular matrix (see methods for more details).

One major goal of this study was to generate resources of value to our lab and the field as a whole. To this end we have generated at least ten cryopreserved stocks from each line. Each stock can be thawed and expanded indefinitely. At least one stock from each individual has been tested and all lines thaw reliably. Furthermore, no lines have shown culture difficulties after thawing. This panel represents the largest stock of characterized non-European iPSCs to date.

Regulatory variation is lower in iPSCs

The faithfulness of iPSCs as a model of embryonic stem cells (ESCs) is still debated; nevertheless, the similarities are evident [139]. This work represents one of the largest collections of iPSCs obtained from healthy individuals. Moreover, to our knowledge this is the only large collection of iPSCs from individuals of African ancestry. Thus, this panel represents a powerful cellular model to study gene regulation at an embryonic-like state. Unique to this study, we have focused on three regulatory phenotypes, mRNA (RNA-seq; n=59), chromatin accessibility (ATAC-seq; n=58), and DNA methylation levels (EPIC array; n=58), to obtain a multi-level understanding of gene regulation in

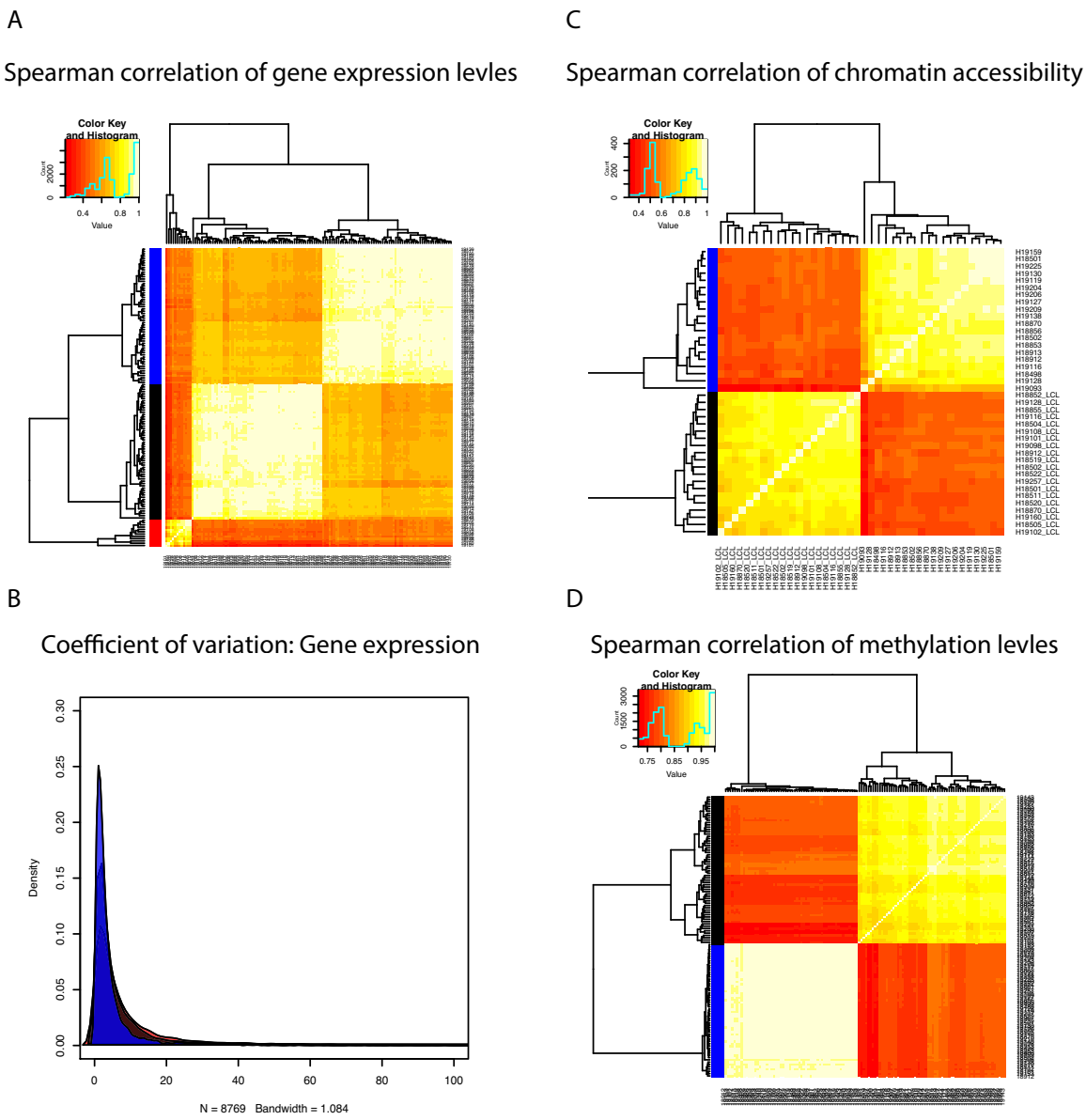

Figure 4.2 Regulatory variation is lower in iPSCs

Heatmaps generated from pairwise spearman correlations of A) gene expression, C) chromatin accessibility and D) DNA methylation levels. Coefficient of variation calculated from gene expression levels in iPSCs, LCLs and iPSC-CMs. In all figures blue denotes iPSCs, black denotes LCLs, and red denotes iPSC-CMs. In all figures iPSCs are the most homogenous.

iPSCs. Furthermore, data from each individual were collected at the same time from the same population of cells (see methods). By using three layers of regulatory data we are able to provide a more comprehensive picture of gene

regulation in iPSCs. To compare gene regulation in iPSCs to other cell types we have differentiated 12 individuals into iPSC-derived cardiomyocytes (iPSC-CMs; see methods) and collected gene expression data (RNA-seq). Additionally, we utilized data previously collected from Yoruba LCLs [36,37,138].

We began our analysis by examining the different trends in overall gene expression between cell types. It became immediately apparent that gene expression in iPSCs is more homogenous than gene expression in LCLs or iPSC-CMs (Fig. 4.2 A&B). This is consistent with a model where embryonic cells are tightly regulated and developmental processes are canalized. We next turned our attention to the methylation and chromatin accessibility data in iPSCs and LCLs. We found that chromatin accessibility had a similar pattern to gene expression and methylation data exhibit an even more striking difference (Fig. 4.2 C&D). While these data may suggest that our study will have lower power to detect genetic associations with gene regulation, recent work from our lab demonstrates that the gene expression variation in iPSCs segregates better by individual than gene expression variation in LCLs [142].

Inter-individual genetic variation drives regulatory differences in iPSCs

After examining overall gene expression patterns we set out to characterize the effect of genetic variation on gene regulation. At a false discovery rate (FDR) of 10%, we have identified thousands of putatively *cis* genetic associations (see methods) with gene expression (eQTLs: 1,629; Figure

4.3A), chromatin accessibility (caQTLs: 2,130), and DNA methylation (meQTLs: 29,782). Although regulatory phenotypes display lower inter-individual variance in iPSCs compared to LCLs, we maintain equal or greater power to detect QTLs when using similar sample size (eQTLs: 1,167; caQTLs: 2,260). Using a recently developed method to identify eQTLs in small sample sizes (see methods) [143] we were able to identify 517 genes where gene expression was associated with at least one genetic variant in iPSC-CMs. This represents the first study to our knowledge that has identified eQTLs in iPSC-derived cell types

Next we set out to characterize the properties of QTLs identified in LCLs and iPSCs. In general we find such properties are well matched across cell types. In particular there appears to be no difference in the average distance between a genetic variant and the associated locus (gene/peak/CpG) across cell types. Some small but significant differences in effect size were identified between cell types; however, these are difficult to interpret and fluctuate depending on the regulatory phenotype. Moreover, when focusing on eQTLs that are significant in both tissues, we see that the effect size of the QTL tracks quite well (Fig 4.3B). These results suggest a high degree of sharing between QTLs and that in general genetic variants affect gene regulation through the same mechanisms regardless of cell types.

Using the pi1 estimate (1-pi0) developed by Story and Tibshirani we estimated the proportion of QTLs that were shared between iPSCs, LCLs, and iPSC-CMs. Rather than making a single estimate based off QTLs identified at an

FDR of 10%, we used a sliding scale to show the distribution of sharing at

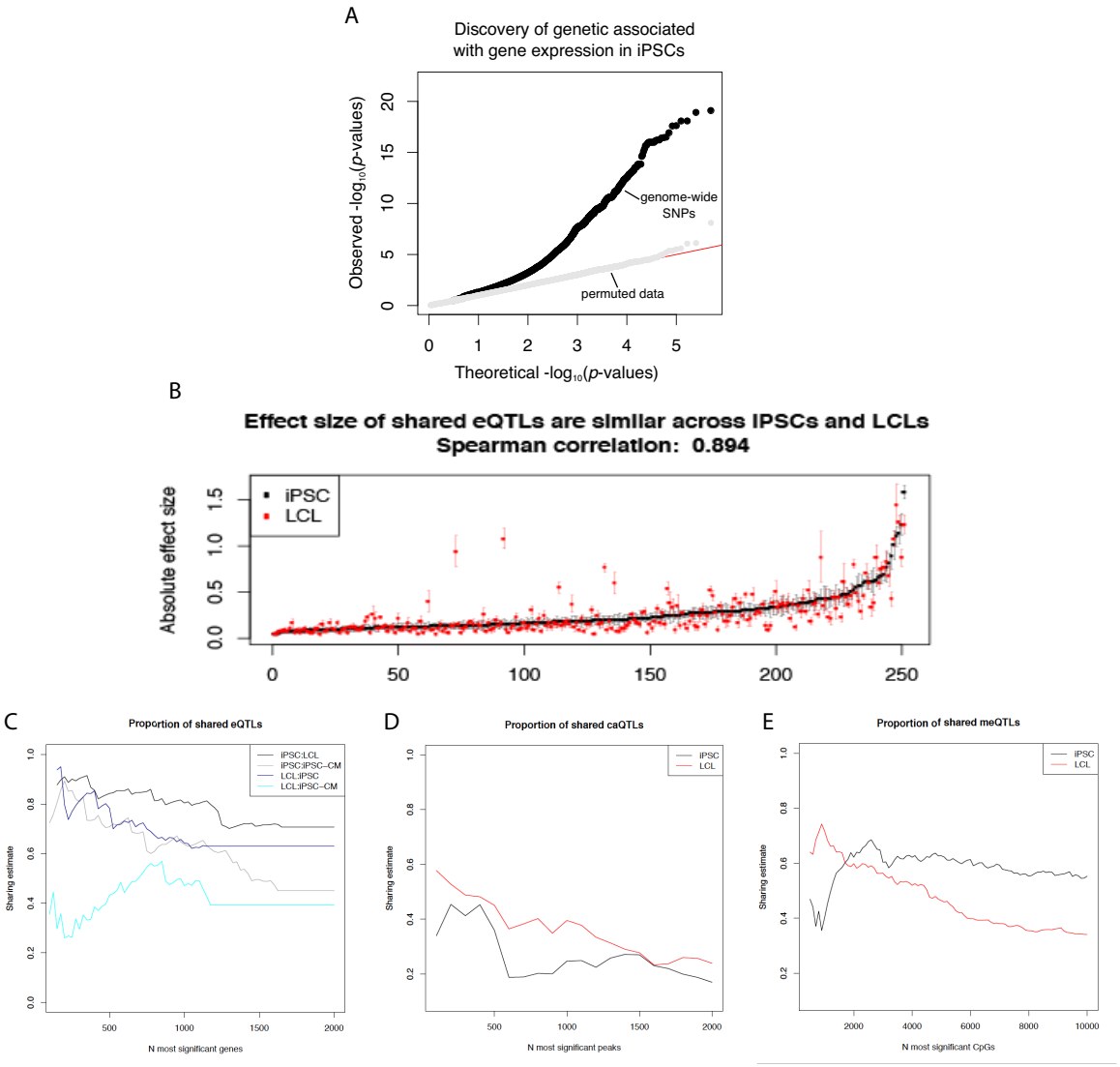

Figure 4.3 Properties of eQTLs across cell types

A) QQ plot of genetic association with gene expression levels. The black dots denote tested SNPs and the grey dots denote permuted data. The red line represents the null expectation. B) Plot of the absolute effect size of eQTLs identified in LCLs and iPSCs. The plot is ordered by the effect size in iPSCs. Standard error of the estimate is plotted around each point. Red points denote LCLs and black points denote iPSCs. C) Estimates of eQTL sharing between iPSCs and LCLs (black lines), iPSCs and iPSC-CMs (grey line), LCLs and iPSCs (dark blue line) and LCLs and iPSC-CMs (light blue line). D-E) Estimates of caQTL and meQTL sharing between iPSCs (black line) and LCLs (red line).

different significance thresholds (Fig. 4.3 C-E). These data demonstrate that gene expression has an extremely high degree of sharing. Indeed, the majority of eQTLs identified in iPSCs are also significant in LCLs (between 71% and 91%; Fig. 4.3C). While the proportion of sharing is lower when considering iPSC-CMs (Fig. 4.3C), this is not unexpected given the difference in sample size. Interestingly, the patterns of sharing differ slightly between iPSCs and LCLs. Namely, eQTLs identified in LCLs exhibit a continual increase in the degree of sharing with iPSCs as the significance threshold increases (Fig. 4.3C). This sharing is maximized when considering only the 150 most significant genes in LCLs (the most stringent threshold). However, eQTLs identified in iPSCs have the largest degree of sharing with LCLs at a slightly more relaxed threshold. These results suggest that there is a higher degree of iPSC specific eQTLs with very low p-values compared to LCLs. Moreover, this pattern is replicated across the regulatory phenotypes tested here (Fig. 4.3 C-E). The proportion of sharing shown here is similar to previous estimates of sharing between iPSCs and somatic eQTLs [141].

Next, we attempted to identify cell type specific eQTLs. Identifying genetic variants with cell types specific effects on gene expression is a difficult task and has been the focus of many previous efforts [32,144]. Here we begin by examining genes that have at least one significant eQTL (eGenes) in iPSCs but were expressed at too low of a level to be tested in LCLs. There are 498 such genes, accounting for 31% of all genes with an eQTL. This is higher than the inverse where only 24% of eGenes in LCLs were not expressed in iPSCs. For

the remaining genes (those tested in both tissues) we used a fairly naïve approach to identify cell type specific eQTLs. First we removed any gene that was significant in both LCLs and iPSCs even if the lead variant differed between the two, and considered only those cases where the lead variant from the cell type where the eQTL was identified was tested in the second cell type. This left us with 533 genes in iPSCs and 530 genes in LCLs. Next we identified cell type specific eQTLs using a two p-value, such that variants significant in one tissue must have a p-value greater than 0.2 in the second tissue. We found nearly identical proportions of genes with a cell type specific eQTL in iPSCs (0.53; n= 285) and LCLs (0.52; n= 278).

One characteristic difference we observed between cell type specific eQTLs and all eQTLs is the distance between the lead variant and the transcription start site (TSS). The median distance in cell type specific eQTLs is significantly larger (iPSC: 35kb, $P < 10^{-3}$; LCL: 35kb, $P < 10^{-5}$) compared to all eQTLs (iPSC: 28kb; LCL: 24kb). This difference is further pronounced (iPSC: 16kb; LCL: 17kb) when focusing on iPSCs that are shared across both cell types (association significant at 10^{-5} in both cell types). This is consistent with a model in which enhancers play a larger role in cell type specific gene regulation. To more explicitly examine this pattern we performed a hierarchical model (see methods) using cell type specific and shared annotations for chromatin states [137], transcription factor binding [145], and caQTLs. We identified enrichments that further suggest cell type specific eQTLs are enriched in enhancers and cell type specific caQTLs (Fig 4.4). These results led us to further examine the effect

of genetic variation on chromatin regulation as a putative mechanism driving the majority of cell type specific QTLs.

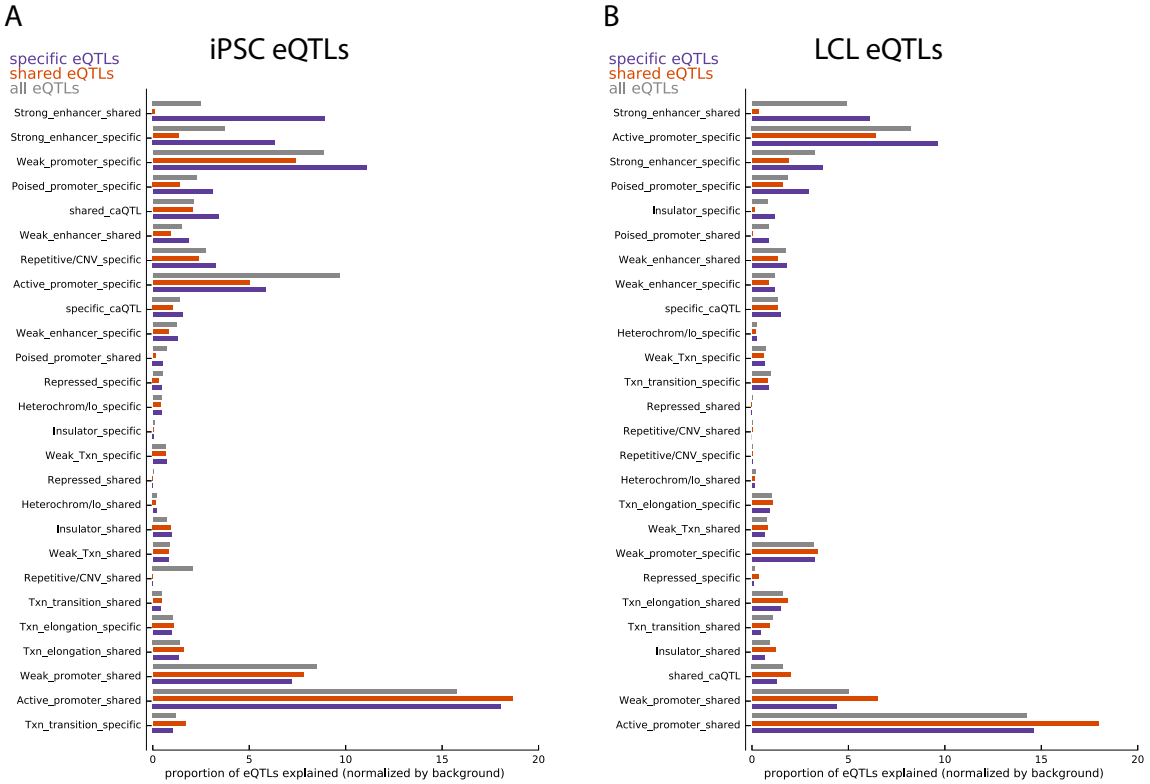

Figure 4.4 Regulatory annotations driving cell type specific and shared eQTLs

Estimates of the proportion of eQTLs explained by chromatin state annotations, genomics annotations, and caQTLs obtained from the hierarchical model in A) iPSCs and B) LCLs. The annotations are separated by shared or specific – i.e. present in both cell types or present in cell type where the eQTL was identified. The bar plots represent eQTLs that are cell type specific (purple), shared (orange), or all eQTLs (grey). These plots are ordered by the difference in the proportion of cell type specific vs shared eQTLs explained such that features explaining more of the cell type specific eQTLs are on top.

Chromatin regulation drives the majority of cell type specific QTLs

We postulated that by focusing on genetic variants affecting chromatin accessibility in a cell type specific manner, we may be better able to dissect the

mechanisms that govern cell type specific eQTLs. For this analysis we limited

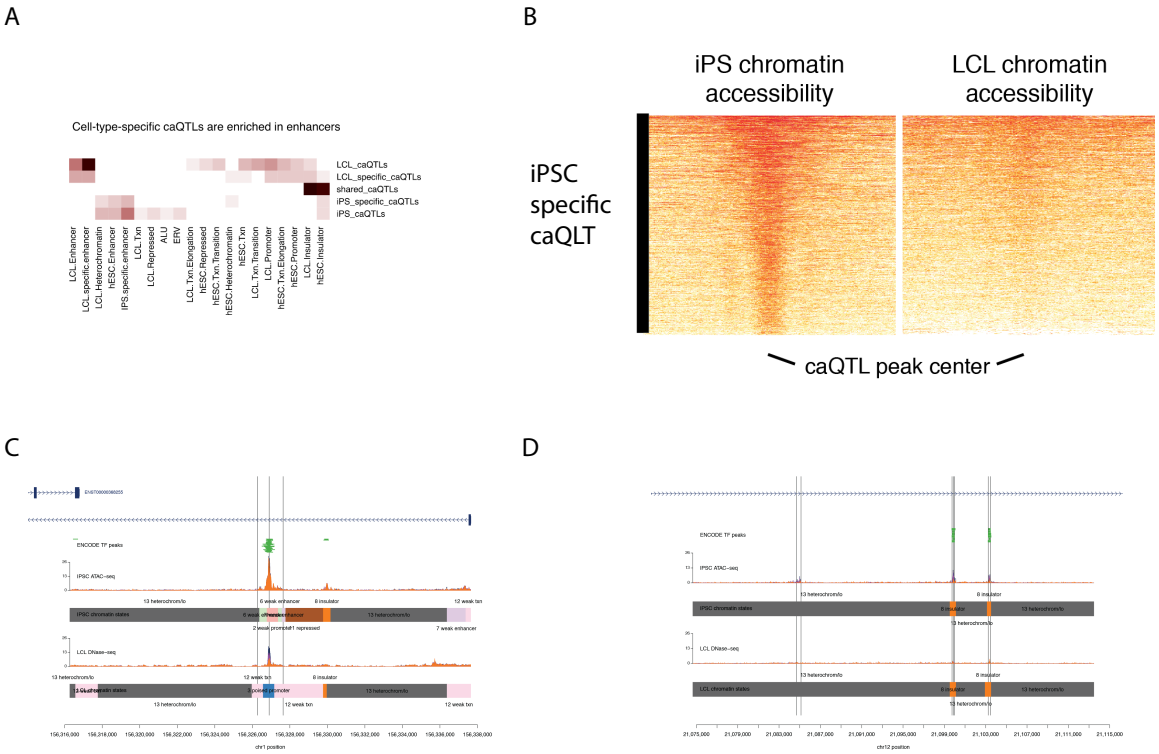

Figure 4.5 cell type specific caQTLs

A) A heatmap showing enrichment of cell type specific and shared caQTLs in regulatory annotations. A darker red denotes a stronger enrichment. B) Heatmap of chromatin accessibility in iPSCs and LCLs at iPSC-specific caQTL peaks. The darker orange denotes increase accessibility. C) An example of an LCL specific caQTL where the peak is accessibility in both regions (model 2) and thus the genetic variant is likely disrupting a cell type specific TFBS. The top row of the plot denotes gene location. The next row shows TF peaks from ChIP-seq data. The density plots show chromatin accessibility levles aggregated by caQTL genotype for iPSCs on the top and LCLs on the bottom. Under the density plots are cell type specific chromatin states. The vertical lines denote the accessibility peak of interest D) An example of an iPSC specific caQTL that is affecting distal regions (model 3). In this particular example the genetic variant is disrupting an insulator element.

our focus on caQTLs where the genetic variants are within a chromatin accessibility peak. Here we examine the chromatin patterns that lead to cell type specific and shared QTLs. Again we identify specific and shared caQTLs using a

two p-value cutoff method (see methods). We first look for enrichment of cell type

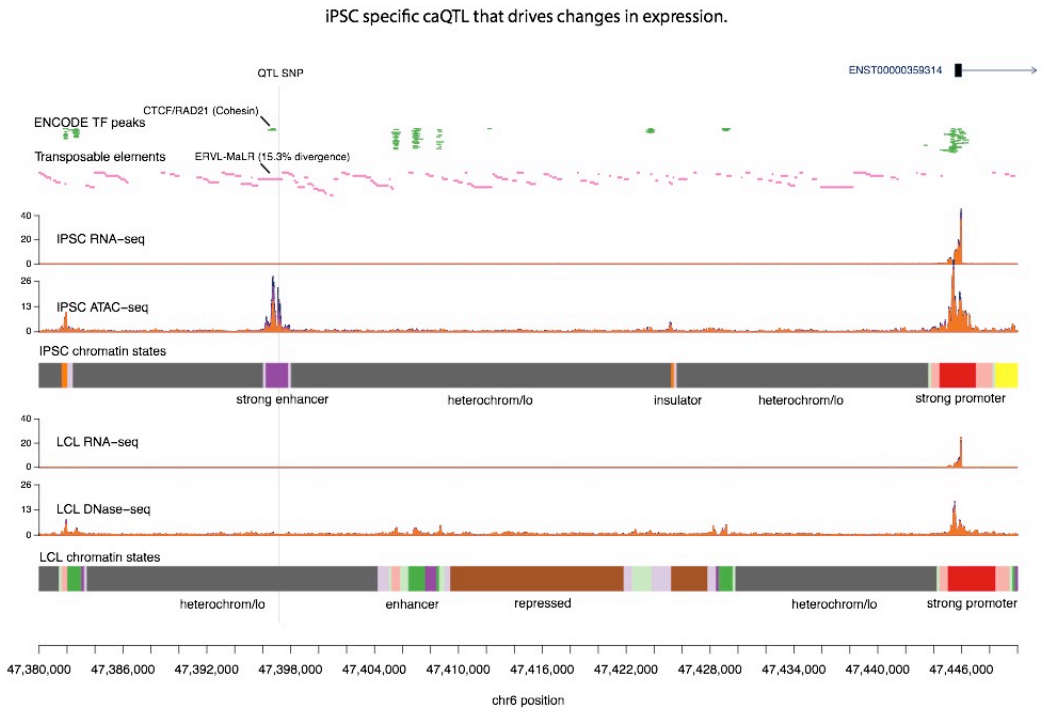

Figure 4.6 An iPSC specific caQTLs that drives cell type specific changes in expression

This example shows an iPSC specific caQTL residing within an iPSC specific chromatin accessibility window that drives cell type specific expression changes. The top row of the plot denotes gene location. The next row shows TF peaks from ChIP-seq data. The third row displays transposable elements. The density plots show gene expression levels and chromatin accessibility levels aggregated by caQTL genotype for iPSCs on the top and LCLs on the bottom. Under the density plots cell type specific chromatin states are displayed. The vertical line denotes the position of the caQTL.

specific caQTLs in chromatin states to confirm the patterns observed in cell type specific eQTLs. Indeed, we see an enrichment of cell type specific caQTLs in cell type specific enhancers (Fig. 4.5 A). In an attempt to identify chromatin patterns that drive cell type specific caQTLs we propose three general models: 1) the region containing the putatively casual SNP is only accessible in one cell type driving the specificity of the QTL 2) the region is accessible in both cell types, but

the SNP is disrupting different transcription factor binding sites 3) the SNP is distal to the region of interest and the change is happening due to interactions between regions. The majority of cell type specific caQTLs fall into the first model (~80%; Fig. 4.5B), a small number of caQTLs are consistent with the second model (~10%; Fig. 4.5C), and a very limited number of examples are of the third model (< 10%; Fig. 4.5D). When we examine shared caQTLs the opposite trend is observed. Namely, we see almost all peaks are accessible in both cell types.

While not all cell type specific caQTLs are also associated with gene expression, 77% of all cell type specific caQTLs that are also eQTLs show a cell type specific pattern (n= 57; $P < 10^{-5}$). One example, shown here, demonstrates a case of the first model, where the caQTL is also an eQTL. Interestingly, the gene affected by this putative enhancer, *CD2AP*, is expressed at similar levels in both cell types but only an eQTL in iPSCs (Fig. 4.6). These results demonstrate how using multiple layers of regulatory data can help us dissect the mechanisms underlying eQTLs.

iPSC-derived cardiomyocytes replicate expression variation in primary hearts

A major goal of this paper is to demonstrate the usability of the iPSC system for the study of complex traits, particularly in hard to collect and disease relevant tissues. To this end we have performed a number of analyses aimed at characterizing the fidelity of iPSC-CMs. Using gene expression data from iPSCs,

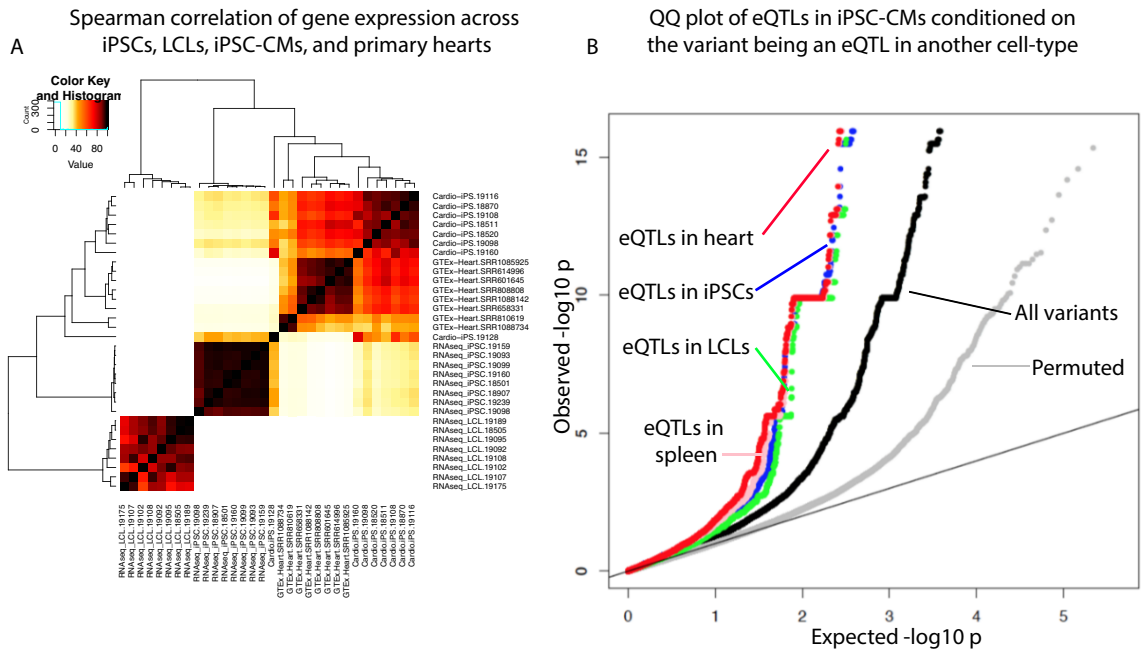

Figure 4.7 iPSC-CMs replicate expression variation from primary heart tissue

A) Spearman correlations of gene expression levels from iPSCs, LCLs, iPSC-CMs, and primary heart tissue. Darker red denotes a stronger correlation. Gene expression levels from iPSC-CMs and primary heart tissues cluster together. B) A QQ-plot of eQTLs identified in iPSC-CMs conditioned on being an eQTL in other cell types. He we examined eQTLs that were significant in hearts (red), spleen (pink), iPSCs (blue), LCLs (green). These are compared to all the variants tested in iPSC-CMs (black) and permuted data (grey). The deviation from the null observed in the permuted data is due to the small sample size. The variants that were eQTLs in primary heart tissue deviate from the line faster than eQTLs from any other tissue. These results suggest that eQTLs in iPSC-CM are most similar to eQTLs identified in primary heart tissue.

iPSC-CMs, and, and tissues collected by the GTEx consortium [32] we performed hierarchical clustering. We find that iPSC-CMs are most similar to the GTEx tissue “Heart Left Ventricle” (Fig 4.7A). Continuing with the GTEx data, we examined the enrichment of eQTLs identified across all tissues in iPSC-CMs. Again, we find eQTLs identified in the “Heart Left Ventricle” are most enriched in iPSC-CMs. As a validation independent of GTEx tissues we used gene ontology enrichment analysis and found eGenes identified in iPSC-CMs are enriched in

many biological processes related to heart function (Table 4.1). Finally, we used a

GO.ID	Term	Annotated	Significant	Expected	classicfisher
GO:0003012	muscle system process	250	25	11.49	0.0002
GO:0002026	regulation of the force of heart contrac...	22	6	1.01	0.00036
GO:0006936	muscle contraction	211	21	9.7	0.00069
GO:0030049	muscle filament sliding	25	6	1.15	0.00076
GO:0033275	actin-myosin filament sliding	25	6	1.15	0.00076
GO:0002704	negative regulation of leukocyte mediate...	17	5	0.78	0.00078
GO:0002698	negative regulation of immune effector p...	45	8	2.07	0.0009
GO:0070252	actin-mediated cell contraction	71	10	3.26	0.00141
GO:0030048	actin filament-based movement	86	11	3.95	0.00185

Table 4.1 Gene ontology enrichment of heart specific processes in iPSC-CM eGenes

polygenic method RolyPoly (see methods) to identify enrichments of GWAS signal in cell type specific gene expression. We examined four GWAS traits:

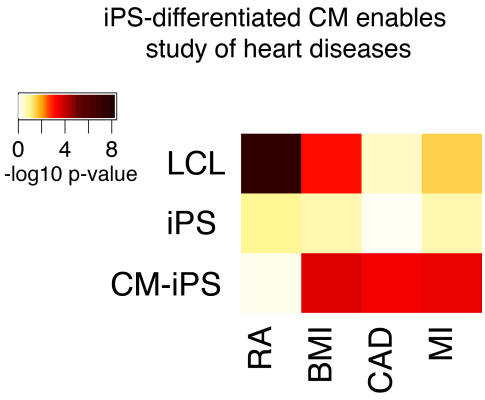

Figure 4.8 iPSC-CMs enable the study of heart disease phenotypes

Enrichment of trait-specific GWAS signal in genes with cell type specific expression. The darker red indices a higher degree of enrichment.

rheumatoid arthritis (RA), body mass index (BMI), coronary artery disease (CAD), and myocardial infarction (MI). Gene expression in iPSC-CMs is enriched for GWAS signal from BMI, CAD, and MI, while LCLs gene expression is enriched for RA and BMI (Fig 4.8). Taken together these results suggest that the gene expression patterns observed in iPSC-CMs replicate those observed in

primary heart tissue, making iPSC-CMs a powerful model in which to study heart specific traits.

4.4 Discussion

Here we have established a unique resource in 59 fully characterized iPSC lines. These lines derived from LCLs obtained from Yoruba individuals originally collected as part of the HapMap project. We believe this resource will be of great value to our lab as well as others. Indeed, we have already had and met requests to share a number of these lines with other labs. We have performed a deep characterization of the genetic architecture underlying inter-individual variation in gene regulation. To our knowledge, this study represents the second largest characterization of gene expression and the largest characterization of chromatin accessibility and DNA methylation in iPSCs [141]. Furthermore, by combining data from LCLs, iPSCs and iPSC-CMs we have for the first time collected multiple regulatory phenotypes in three cell types from the same panel of individuals.

We have identified novel QTLs in two cell types (iPSCs and iPSC-CMs). We show here that the reduced variation in regulatory phenotypes found in iPSCs does not diminish our ability to identify QTLs. We have identified a list of iPSC specific and LCL specific eQTLs. These eQTLs allowed us to identify chromatin features that drive cell type specific and shared eQTLs. The enrichments we observed suggested that genetic variants within enhancers

driving changes in chromatin at loci distal to the TSS were the major drivers of cell type specific eQTLs. This finding is consistent with what has been reported previously [32] and is supported by a large body of work demonstrating the tissue specificity of enhancers [18,67,146-150].

The results presented here significantly advance our knowledge of cell type specific eQTLs. Others have reported and characterized the genomic architecture of cell type specific eQTLs [32,144], yet this is the first study using additional regulatory phenotypes to identify putative mechanisms driving such eQTLs. In particular, the chromatin accessibility data presented here allowed us to identify cell type specific caQTLs within enhancer elements that have a cell type specific effect on expression. These results provide a definite mechanism by which cell type specific eQTLs can act.

Finally, we show that iPSC-CMs are a useful model for studying heart specific traits. Gene expression patterns in iPSC-CMs replicate those of primary heart tissue and genetic variation has similar effects. These results when taken together with other recent work [117] suggest that iPSC-CMs are a powerful model in which to study heart specific phenotypes. Importantly, this highlights the power of the iPSC system as a whole. Future studies using this panel of iPSCs will be able to assay dynamic gene expression by characterizing gene expression during differentiation, in multiple cell types from the same individuals, and in terminally differentiated cell types subjected to environmental perturbations. The study of dynamic gene regulation in these model, in conjunction with newly developed genome editing technologies [151] will allow

researchers to functionally follow up on putatively causative alleles. The research presented here is a valuable first step towards that goal.

4.5 Materials and Methods

iPSC generation

We reprogrammed LCLs into iPSCs using an episomal reprogramming approach described previously [121,140]. Briefly, we transfected 1 million LCLs (Amaxa™ Nucleofector™ Technology; Lonza) with 1ug of oriP/EBNA1 PCXLE based episomal plasmids that contain the genes OCT3/4, SOX2, KLF4, L-MYC, LIN28, and an shRNA against p53. Cells were cultured in suspension for seven days after transfection in hESC media (DMEM/F12 supplemented with 20% KOSR, 0.1mM NEAA, 2mM GlutaMAX, 1% Pen/Strep, 0.1% 2-Mercaptoethanol, 25ng/ul of bFGF and .5mM NaB). On the 8^{th} day we plated a range of 8,000 - 32,000 transfected cells per well in a 6-well plate coated with gelatin and seeded with irradiated CF1 mouse embryonic fibroblasts (MEF). Four days after the initial plating NaB was removed from the hESC media. Within 21 days colonies were visible and manually passaged onto a freshly prepared gelatin plate MEF. Manually passaging continued weekly for ten weeks. After ten passages of growth cells were expanded and at least ten stocks of cells were cryopreserved. Colonies that were not cryopreserved were then transitioned to feeder-free conditions and cultured for at least an additional three passages before collecting cell pellets for analysis. Feeder-free cultures were grown using 0.01mg/cm2 (1:100) hESC-grade Matrigel and Essential 8 (E8) media. Feeder-free passaging

is enzymatic rather than manual and was performed using DPBS supplemented with 0.5mM EDTA.

iPSC characterization

All iPSC lines were characterized for pluripotency and stability using three methods. First, we confirmed the ability of lines to differentiate to all three germ layers using the embroyid body (EB) assay. Lines were manually dissociated from their culture dish in large pieces. This material was then cultured in a suspension plate using the hESC media described above without bFGF for one week, while dense spherical EBs form. EBs are then plated into 12 well plates with gelatin and cultured in EB medium (DMEM supplemented with 10% FBS, 0.1mM NEAA, 2mM GlutaMAX) for one week. EBs in each well were then immunostained for cell types from all three germ layers (Fig 4.1A). Next, all lines were karyotyped to search for large genomic rearrangements (Fig. 4.1C). Lines were karyotypes by the WiCell Research Institute (Madison, WI). Only one line, 19128, showed large genomic rearrangements that were not known rearrangements segregating in the population. The rearrangement observed in this line is a hallmark rearrangement of follicular lymphoma and thus was likely present in LCLs rather than a result of the reprogramming process. Finally, a classifier, PluriTest [130] was applied to gene expression data (Illumina HumanHT-12 array) to assay pluripotency bioinformatically. The classifier compares gene expression levels from uncharacterized lines to a “gold standard” panel of embryonic stem cells and iPSCs. Two metrics are obtained from this

method, a pluripotency score and a novelty score. The pluripotency score represents goodness of fit of canonical pluripotency genes in the sample. The novelty score represents the deviance of non-pluripotency genes in the sample. All of the lines here pass the suggested empirical threshold (Fig. 4.1B).

iPSC-derived cardiomyocyte differentiation

Differentiation from iPSCs to cardiomyocytes was done using slight modifications of existing protocols [152,153]. iPSCs cultured in feeder-free conditions cells were seeded to a 10cm dish three to five days prior to differentiation. When cells were 70-100% confluent (i.e. the total amount of dish occupied by cells) E8 media was replaced with heart media (RPMI supplemented with B27 minus insulin, 2mM GlutaMAX, and 100mg/mL Pen Strep) with the addition of 1:100 matrigel and 12uM of the GSK-3 inhibitor CHIR which activates WNT signaling (day 0) [152]. After 24 hours media was replaced with new heart media (day 1). After an additional 48 hours media was replaced with new heart media with the addition of 2uM of the WNT inhibitor WntC59 [152]. (day 3). Cells were cultured in the media with WntC59 for 48 hours. The cells were then cultured in heart media with regular media changes until day 14. Clusters of spontaneously beating cells were typically visible between 7 and 12 days. On day 14 heart media was replaced with CDM3 with lactate (RPMI without glucose, 75 mg/ml human albumin, 213 ug/ml L-ascorbic acid 2-phoshate, 5mM sodium DL-lactate, and 100mg/mL Pen Strep) . CDM3 with lactate preforms a metabolic purification. Namely, the majority of cells cannot use lactate as their primary

source of energy, leaving a culture significantly enriched for cardiomyocytes [153]. Every other day media changes were performed until day 20. By day 20 the cells had generally formed into large three-dimensional sheets of beating cells. To make a more uniform sheet of cells we dissociated the cultures using 0.05% trypsin and replated cells into six well plates at a density of 1.5 million cells per well. Cells were then cultured in galactose media (DMEM without glucose, 1.7 mg/mL galactose, 1mM Na pyruvate, 5mM HEPES, 2mM GlutaMax, 10% FBS, and 100mg/mL Pen Strep). The galactose-based media helped to mature cardiomyocytes by forcing aerobic metabolism [52,154]. Regular media changes with galactose media continued for the duration of the experiment. After an additional four days (day 25) cells were moved to an incubator at physiological oxygen levels (10%). Five days after cells had been moved to physiological oxygen levels (day 29) they were subjected to electrical stimulation for three days to help further mature the cells [155] and standardize beating rate across wells and lines.

Sample Collection

After at least three passages in feeder-free conditions iPSCs were passaged into a 10cm culture dish. At near full confluence cells were enzymatically dissociated and counted. After dissociation further collection is done on ice or in a temperature controlled centrifuge. One 10cm dish yields between 3 million and 15 million cells. From each line 400,000 cells were divided into two tubes to be used for ATAC-seq [156]. The tagmentation step of the

ATAC-seq protocol was performed immediately on the two cell pellets containing 200,000 cells each. The library preparation of ATAC-seq samples was done in larger batches at a later time. The remaining material was split between three tubes for RNA and DNA extractions. We isolated RNA and DNA using the Zymo dual extraction kits (Zymo Research) with a DNase treatment during RNA extraction (Qiagen) on a single cell pellet from each line. 50 bp single-end RNA sequencing libraries were generated from extracted RNA using the Illumina TruSeq kit as directed by the manufacturer. ATAC-seq and RNA-seq was performed on an Illumina 2500. Extracted DNA was bisulphite-converted and hybridized to the Infinium MethylationEPIC array (Illumina) at the University of Chicago Functional Genomics facility.

iPSC-CMs were collected on ice using manual dissociation. One pellet was collected from each well of the six wells (see above). Generally between one and three wells were obtained per individual. We isolated RNA and DNA using the Zymo dual extraction kits (Zymo Research) with a DNase treatment during RNA extraction (Qiagen) on a single cell pellet from each line. 50 bp single-end RNA sequencing libraries were generated from extracted RNA using the Illumina TruSeq kit as directed by the manufacturer. RNA-seq was performed on an Illumina 2500.

RNA-seq processing

RNA-seq from LCLs [30] and iPSCs were mapped using the STAR RNA-seq aligner standard settings. RNA-seq reads from cardiomyocytes were

mapped using Subread allowing for two mismatches. Reads overlapping SNPs were remapped to reduce reference bias as described previously [143]. Only reads with a MAPQ greater than ten were retained.

ATAC-seq processing

Paired end ATAC-seq reads were mapped using bowtie2 allowing for two mismatches per read. The ATAC-seq protocol works by randomly inserting sequencing adapters into open chromatin via a tagmentation enzyme. One unfortunate side effect of this procedure is an extreme enrichment of reads originating from mitochondrial reads (between 25%-75% of reads). Only nuclear reads are maintained for analysis. After mitochondrial reads are removed we remove all duplicate fragments (duplicates of both read pairs) and reads with a MAPQ less than ten. Each mate represents an independent tagmentation event and therefore after mapping and duplicate removal reads are treated as single end in all future analyses.

DNase processing

Previously collected DNase-seq from LCLs was used to assay chromatin accessibility. Reads were mapped using a custom mapper, which has been previously described in depth [97]. In this study counts per base directly obtained from a previous study were used [37].

Methylation array processing

Methylation levels were assayed using the Infinium MethylationEPIC array (Illumina) in iPSCs and the Infinium HumanMethylation450 array (Illumina) in LCLs. Methylation data from LCLs were obtained from a previous study [36]. In iPSCs a number of steps were taken to ensure high quality data. First, to enable accurate quantification of methylation levels all probes that contained a SNP with a MAF greater than 5% in the population were removed. Next, we removed all CpGs that were not detected in 75% of individuals. CpGs on the X or Y chromosome were removed.

Identifying eQTLs

To identify eQTLs in iPSCs and LCLs we fit expression levels to a standard normal within each individual (iPSC: n= 59, LCL: n= 59). We also accounted for unknown confounders by removing principal components from the LCL data. Genotypes were obtained using impute2 as described previously [31]. As in previous work we are limited to examining putatively *cis* acting genetic variants. Therefore, we only consider variants within 50kb of genes. To identify association between genotype and gene expression we used the fastqtl software [157]. This program performs a linear regression between the genotype of a genetic variant and expression level. After the initial regression a variable number of permutations are performed to obtain a gene-wise adjusted p-value [157]. To identify significant eQTLs we use Story's q-value [109] on the adjusted p-values. Genes with a q-value less than 0.1 are considered significant.

The sample size of iPSC-CMs in this study was prohibitive to call eQTLs using a standard regression model. We therefore utilized the combined haplotype test (CHT) [143] to identify eQTLs. This method allows one to identify eQTLs with small sample sizes by using both regression and allelic imbalance tests in combination. Here we focus on variants within 25kb of a gene. Following the procedure outlined by the authors [109] we performed the CHT and one permutation of the CHT. Given the small sample size the test is not well calibrated, showing significant signal in the permuted version of the test. At the suggestion of the authors we identified significant SNPs by performing Story's q-value correction [109] on the null data. We then identified the largest p-value in the null data with a q-value less than 0.1. We used this p-value as a threshold in the non-permuted data to identify significant eQTLs.

Identifying meQTLs

To identify meQTLs in iPSCs and LCLs we fit methylation levels to a standard normal within each individual (iPSC: n= 58; LCL: n= 64) and unknown confounders are accounted for by removing principal components from the data (iPSC: 6 PCs removed; LCLs: 5 PCs removed). In accordance with previous work, genetic variants within 3kb of a CpG were tested for associations with methylation levels. meQTLs were identified using the fastqtl software following the procedure described above. We inherently identified a larger number of meQTLs in iPSCs compared to LCLs due to the increase in the number of CpGs tested. However, we also compared only the CpGs shared across both arrays

and found that we were still able to identify more meQTLs in iPSCs (n= 7,958; n= 5,738).

Identifying caQTLs

We began by identifying a set of chromatin accessibility peaks that were shared in both iPSCs and LCLs. Of note, the chromatin accessibility data in iPSCs is from ATAC-seq while the chromatin accessibility data in LCLs is from DNase-seq. Chromatin accessibility levels were fit to a standard normal within each individual (iPSC: n= 55; LCL: n= 68) and principal components were removed to account for unknown confounders (iPSCs: 2 PCs removed; LCLs: 4 PCs removed). Associations between genetic variants within 25kb of a peak and chromatin accessibility levels were identified using a linear regression. To obtain a locus-wise adjusted p-value the individual labels of genetic variants for each peak were shuffled and the regression was re-run. This permutation was performed 100,000 times and the adjusted p-value is the number of times a p-value from the permutation was lower than the original lowest p-value divided by 100,000. Story's q-value [109] was applied to the adjusted p-values and a locus was considered significant if the q-value was less than 0.1.

Estimating QTL sharing

Story and Tibshirani developed a method to estimate the true proportion of null statistics from a given p-value distribution [109]. This metric (pi0) can be used to calculate the proportion of significant tests from a p-value distribution by

taking 1 - pi0 (pi1). Here we calculate pi1 for eQTLs, caQTL, and meQTLs between cell types. To obtain a better estimate of the true sharing we generated pi1 statistics for a range of stringencies. Specifically, for eQTLs and caQTLs we calculated pi1 cumulatively from the top 150 most significant genes/loci to the top 2000 most significant genes/loci in intervals of 25 genes/lcoi. For meQTLs we calculated pi1 from the top 500 CpGs to the top 10,000 CpGs in intervals of 100 CpGs. As is clear from the density plots (Fig. 4.3C-E), small deviations in threshold choice can create local valleys and peaks in sharing estimates. This method allows us to see sharing across a wide space of stringencies.

Identifying specific and shared eQTLs

We first removed loci that were tested in only one cell type. Next, any locus with a significant association (even with a different lead variant) in both cell types was removed. A QTL was considered cell type specific if significant at an FDR of 10% in one cell type and a nominal p-value of greater than 0.2 in the second cell type. QTLs were considered shared if they were significant with a p-value of less than 10^{-5} in both cell types.

GO term enrichment analysis

GO terms were identified using the bioconductor package topGO [158]in R. Genes that had at least one significant association with a genetic variant were compared against a background of all genes tested. Only the ontology terms

associated with “biological processes (BP)” were considered. Fisher’s exact test was used to generate p-values.

Hierarchical model

The hierarchical model used here was developed to identify causal SNPs from in eQTLs studies by incorporating annotations such as chromatin states or chromatin accessibility. The method is explained in detail elsewhere [31] and the software used to implement the model is available here: https://github.com/rajanil/qtlBHM. For the purposes of this paper we sought to identify annotations that were informative in cell type specific eQTLs when compared with shared eQTLs and all eQTLs.

GTEx data

Only summary statistics were collected from the GTEx data [32]. Specifically, for every gene tested in a tissue, the p-value of the lead variant was obtained. To overlap with eQTLs identified in iPSC-CMs the variant identified in the GTEx data was tested in iPSC-CMs. The QQ-plot was generated from a limited number of tissues for clarity (Fig. 4.7B).

GWAS signal enrichments in gene expression data

RolyPoly is a highly polygenic method that identifies trait-involved cell types by analyzing the enrichment of GWAS signal in cell type specific gene expression genome-wide. First, for each gene we calculate trait association scores by

aggregating GWAS summary statistics from a window (10kb) centered on the TSS. Then, we estimate the individual contribution of each cell type to the observed gene score variance using a generalized linear regression model with normalized gene expression features. For each cell type we estimate an effect size coefficient and standard error, which we use for hypothesis testing. We implemented the RolyPoly method in the rolypoly R package

CHAPTER 5: DISCUSSION

In chapter two I, in collaboration with Xun Lan, identified nearly 14 thousand CpGs whose methylation levels are associated with genetic variation in LCLs. While these results were similar to other recent meQTL studies [34], we leveraged the plethora of information previously collected from these LCLs. Specifically, we identified associations between meQTLs and multiple histone modifications, Pol II binding, DNase I hypersensitivity, and expression. One interesting result that arouse in this study is that meQTLs, which are also eQTLs, often have the same direction of effect in both phenotypes. This challenges the general narrative that DNA methylation is negatively correlated with expression. In these cases the CpG and TSS are almost always quite distant from one another. This suggests that methylation is acting on a non-promoter element and that DNA methylation may context specific effects. These results add to a growing body of evidence that a single genetic variant is often associated with coordinated changes in multiple regulatory phenotypes and further demonstrate the complexity of interactions between such regulatory phenotypes.

In this study we demonstrate that changes in methylation driven by genetic variation often act through disrupting transcription factor binding sites (TFBS). Specifically, changing the binding of transcription factors often affects the methylation levels of CpGs near a given TFBS. In particular, five transcription factors, *CTCF*, *PAX9*, *ESE1*, *STAT5*, and *ZNF274,* have a larger than expected effect on DNA methylation. One drawback of correlative studies such as this is that it is difficult to identify the order of events – i.e. we do not know what the first

step in the regulatory cascade is. However, by focusing on genetic variants disrupting TFBS we identify a putative mechanism and are likely observing the first step leading to a change in DNA methylation. In other words, this approach suggests that changes in TF binding are frequently a key early step in the regulatory cascade that leads to concerted changes in multiple mechanisms.

In chapter three I, in collaboration with Courtney Burrows, turn my attention to iPSCs. While this model has been used for 10 years some serious questions remain about the usefulness of these cells to study human phenotypes and as a tool for regenerative medicine [61-63]. One major concern was that lingering “epigenetic memory” of the somatic tissue of origin remained after reprogramming [62,112-118]. These previous studies found that when clustering methylation and expression profiles of iPSCs derived from different cell types, the iPSCs would cluster by their cell type of origin [112,114-116]. However, all of the previous work, sans one study examining only gene expression [123], used study designs that confounded inter-individual differences with somatic cell type of origin. These studies were therefore less than ideal to study this phenomenon.

Here we developed an effective study design to examine gene expression and DNA methylation levels in iPSCs derived from two cell types (LCLs and fibroblasts) in four individuals (two males and two females). When comparing DNA methylation and gene expression levels of iPSCs derived from different cell types we see almost no differences. Indeed, we identified only 197 CpGs out of over 300 thousand tested that were differentially methylated between fibroblast-derived iPSCs (F-iPSC) and LCL-derived iPSCs (L-iPSCs), and only 37 of these

were near a gene. Even more shockingly, we identified only one differentially expressed gene between F-iPSCs and L-iPSCs. In an effort to measure the contribution of both individual and cell type of origin more explicitly, we employed a linear mixed model. Using this model we demonstrate that individual accounts for the majority of observable variation in both gene expression and DNA methylation. An additional study, which came out after the publication of the work presented here, validated our findings in an additional panel of iPSCs [159]. Importantly, Kyttala et al. also differentiated their iPSCs derived from two muscle and blood cells into iPSC-derived blood cells. Their results suggest that individual genetic variation is the largest contributor to variation in the iPSC-derived tissues [159].

The results presented in chapter thre challenge the commonly held belief that “epigenetic memory” is one of the largest drivers of regulatory differences in iPSCs. Furthermore, results presented here and elsewhere [159] demonstrate that iPSCs have gene expression and DNA methylation patterns are driven by genetic variation. Taken together these results suggest that iPSCs are a suitable cell type in which to study inter-individual variation in gene regulation.

Finally in chapter four I, in collaboration with Yang Li and Anil Raj, established and characterized a large panel of iPSCs from 59 West African Yoruba individuals. The resource developed here, and the subsequent results, represent four years of work. This panel of iPSCs has a number of unique features. First, the LCLs from which these lines were derived have been extensively characterized for numerous regulatory phenotypes [21,24,26,31,35-

37,40,160]. Next, this is one of the largest collections of iPSCs derived from health individuals, and is, to our knowledge, the largest collection of iPSCs from individuals of African descent. Finally, this is the only panel of iPSCs, to our knowledge, where data on chromatin accessibility, DNA methylation, and gene expression levels has been collected. Moreover, this is the only data set of sufficient size to investigate inter-individual variation where three regulatory phenotypes have been collected at one time and processed in parallel from any cell type.

Using the data generated here we are able to identify thousands of genetic associations with gene expression, chromatin accessibility, and DNA methylation in iPSCs. One additional study has identified eQTLs in iPSCs [141], yet this is the first study to identify meQTLs and caQTLs in iPSCs. After identifying eQTLs in both LCLs and iPSCs, we set out to find cell type specific and shared eQTLs. Similar to other studies examining cell type specific eQTLs we find such eQTLs enriched in enhancer elements, TFBS of transcription factors with tissue specific expression, and cell type specific caQTLs [144]; [32]. These results led us to focus on chromatin accessibility, particularly at distal enhancers, as a putative mechanism underlying cell type specific eQTLs. We were able to identify nearly 350 cell type specific caQTLs in both iPSCs and LCLs. Cell type specific caQTLs generally fall into three models: 1) the region containing the putatively casual SNP is only accessible in one cell type driving the specificity of the QTL 2) the region is accessible in both cell types, but the SNP is disrupting different transcription factor binding sites 3) the SNP is distal to

the region of interest and the change is happening due to interactions between regions. Over 80% of the cell type specific caQTLs we identified here are consistent with the first model. Not all cell type specific caQTLs are also eQTLs, yet, of those over 75% have cell type specific effects. Taken together results suggest that the majority of cell type specific caQTLs and eQTLs reside within enhancer elements active in the cell type of interest. While this result is not unexpected, these results could not have been obtained without characterizing both gene expression and chromatin accessibility.

Finally, we differentiated iPSCs from 12 individuals into iPSC-derived cardiomyocytes (iPSC-CMs). We set out to demonstrate that iPSC-CMs are a viable model for the study of heart specific traits. To this end, we generated gene expression data and compared gene expression levels from iPSC-CMs to gene expression levels in primary tissue collected by the GTEx consortium [32]. Indeed, the iPSC-CMs cluster most similarly to primary heart tissue. Additionally, we found that gene expression in iPSC-CMs captures cell type specific enrichment of GWAS signals. Finally, we identified eQTLs in iPSC-CMs using the combined haplotype method [143]. eGenes identified in this analysis are enriched for biological processes related to heart function. We also found that eQTLs identified here are most enriched for eQTLs identified in primary heart tissue by the GTEx consortium [32]. Taken together we believe these results clearly establish the fidelity of iPSC-CMs and their usefulness to study heart specific traits. These results are bolstered by recent work showing iPSC-CMs

recapitulate doxorubicin-induced cardiotoxicity in a number of breast cancer patients [55].

The work presented in this thesis represents a major advance in our understanding of the mechanisms underlying regulation in gene expression. Additionally, this work has established an iPSC bank that can be used in future research indefinitely. Importantly, the work presented in this thesis makes a beginning at exploring dynamic gene regulation – i.e. gene expression in three cell types representing different developmental stages from the same individuals. The iPSCs generated here are already being used in large-scale studies to explore dynamic gene expression during differentiation and in response to environmental stimulus. Studies of dynamic gene expression in combination with recently developed gene editing techniques [151] promise to usher in a new era of genomics where true dissection and validation of mechanisms underlying inter-individual variation in gene expression is possible.

References

1. BLAKESLEE AF (1914) CORN AND MEN: The Interacting Influence of Heredity and Environment—Movements for Betterment of Men, or Corn, or Any Other Living Thing, One-sided Unless They Take Both Factors into Account. Journal of Heredity 5: 511-518.

2. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-331.

3. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24.

4. Manoel RO, Freitas ML, Tambarussi EV, Cambuim J, Moraes ML, et al. (2015) Mendelian inheritance, genetic linkage, and genotypic disequilibrium at microsatellite loci in Genipa americana L. (Rubiaceae). Genet Mol Res 14: 8161-8169.

5. Dueker ND, Pericak-Vance MA (2014) Analysis of genetic linkage data for Mendelian traits. Curr Protoc Hum Genet 83: 1 4 1-31.

6. Bucci G, Menozzi P (2002) Spatial autocorrelation and linkage of Mendelian RAPD markers in a population of Picea abies Karst. Mol Ecol 11: 305-315.

7. Craig HD, Gunel M, Cepeda O, Johnson EW, Ptacek L, et al. (1998) Multilocus linkage identifies two new loci for a mendelian form of stroke, cerebral cavernous malformation, at 7p15-13 and 3q25.2-27. Hum Mol Genet 7: 1851-1858.

8. Hill ME, Davies KE, Harper P, Williamson R (1982) The Mendelian inheritance of a human X chromosome-specific DNA sequence polymorphism and its use in linkage studies of genetic disease. Hum Genet 60: 222-226.

9. Bush WS, Haines J (2010) Overview of linkage analysis in complex traits. Curr Protoc Hum Genet Chapter 1: Unit 1 9 1-18.

10. Ma J, Daw EW, Amos CI (2010) Power of competing strategies of linkage analysis for complex traits. Hum Hered 70: 55-62.

11. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

12. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197-206.

13. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108.

14. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.

15. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8: e1002822.

16. Hindorff L, MacArthur J, Morales J, Junkins H, Hall P, et al. (2013) A Catalog of Published Genome-Wide Association Studies. Avaiable at: http://www.genome.gov/gwastudies.

17. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22: 1748-1759.

18. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, et al. (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507: 371-375.

19. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365-1369.

20. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-Wide Associations of Gene Expression Variation in Humans. PLoS Genetics 1: e78.

21. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768-772.

22. Stranger BE, Nica AC, Forrest MS, Beazley C, Ingle CE, et al. (2007) Population genomics of human gene expression. Nature Genetics 39: 1217-1224.

23. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, et al. (2013) DNA methylation contributes to natural human variation. Genome Research 23: 1363-1372.

24. Veyrieras J-B, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK (2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genetics 4: e1000214.

25. Brown CD, Mangravite LM, Engelhardt BE (2012) Integrative modeling of eQTLs and cis-regulatory elements suggest mechanisms underlying cell type specificity of eQTLs. arXivorg.

26. Gaffney DJ, Veyrieras J-B, Degner JF, Pique-Regi R, Pai AA, et al. (2012) Dissecting the regulatory architecture of gene expression QTLs. Genome Biology 13: R7.

27. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al. (2010) Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. PLoS Genetics 6: e1000952.

28. Pickrell JK (2014) Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet 94: 559-573.

29. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43: 246-252.

30. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.

31. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, et al. (2016) RNA splicing is a primary link between genetic variation and disease. Science 352: 600-604.

32. Consortium GT (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.

33. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24: 14-24.

34. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, et al. (2013) Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife 2: e00523.

35. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, et al. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology 12: R10.

36. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, et al. (2014) Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLoS Genet 10: e1004663.

37. Degner JF, Pai AA, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390-394.

38. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, et al. (2013) Extensive variation in chromatin states across humans. Science 342: 750-752.

39. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, et al. (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science 342: 744-747.

40. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, et al. (2013) Identification of genetic variants that affect histone modifications in human cells. Science 342: 747-749.

41. Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, et al. (2013) Effect of natural genetic variation on enhancer selection and function. Nature 503: 487-492.

42. Cusanovich DA, Billstrand C, Zhou X, Chavarria C, De Leon S, et al. (2012) The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. Hum Mol Genet 21: 2111-2123.

43. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell 131: 861-872.

44. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science 318: 1917-1920.

45. Yamanaka S, Takahashi K (2006) Induction of pluripotent stem cells from mouse fibroblast cultures. Tanpakushitsu Kakusan Koso 51: 2346-2351.

46. Okita K, Ichisaka T, Yamanaka S (2007) Generation of germline-competent induced pluripotent stem cells. Nature 448: 313-317.

47. Park I-H, Arora N, Huo H, Maherali N, Ahfeldt T, et al. (2008) Disease-specific induced pluripotent stem cells. Cell 134: 877-886.

48. Narsinh K, Narsinh KH, Wu JC (2011) Derivation of Induced Pluripotent Stem Cells for Human Disease Modeling. Circ Res 108: 1146-1156.

49. Josowitz R, Carvajal-Vergara X, Lemischka IR, Gelb BD (2011) Induced pluripotent stem cell-derived cardiomyocytes as models for genetic cardiovascular disorders. Curr Opin Cardiol 26: 223-229.

50. Chang WY, Garcha K, Manias JL, Stanford WL (2012) Deciphering the complexities of human diseases and disorders by coupling induced-pluripotent stem cells and systems genetics. Wiley Interdisciplinary Reviews Systems Biology and Medicine 4: 339-350.

51. Simone C, Nizzardo M, Rizzo F, Ruggieri M, Riboldi G, et al. (2014) iPSC-Derived neural stem cells act via kinase inhibition to exert neuroprotective effects in spinal muscular atrophy with respiratory distress type 1. Stem Cell Reports 3: 297-311.

52. Wang G, McCain ML, Yang L, He A, Pasqualini FS, et al. (2014) Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. Nat Med 20: 616-623.

53. Xie YZ, Zhang RX (2015) Neurodegenerative diseases in a dish: the promise of iPSC technology in disease modeling and therapeutic discovery. Neurol Sci 36: 21-27.

54. Bedut S, Seminatore-Nole C, Lamamy V, Caignard S, Boutin JA, et al. (2016) High-throughput drug profiling with voltage- and calcium-sensitive fluorescent probes in human iPSC-derived cardiomyocytes. Am J Physiol Heart Circ Physiol 311: H44-53.

55. Burridge PW, Li YF, Matsa E, Wu H, Ong SG, et al. (2016) Human induced pluripotent stem cell-derived cardiomyocytes recapitulate the predilection of breast cancer patients to doxorubicin-induced cardiotoxicity. Nat Med 22: 547-556.

56. Chaudhari P, Prasad N, Tian L, Jang YY (2016) Determination of Functional Activity of Human iPSC-Derived Hepatocytes by Measurement of CYP Metabolism. Methods Mol Biol 1357: 383-394.

57. Crunkhorn S (2016) Diabetes: Human iPSC-derived beta-like cells rescue diabetic mice. Nat Rev Drug Discov 15: 382-383.

58. Ou D, Wang Q, Huang Y, Zeng D, Wei T, et al. (2016) Co-culture with neonatal cardiomyocytes enhances the proliferation of iPSC-derived cardiomyocytes via FAK/JNK signaling. BMC Dev Biol 16: 11.

59. Zhou S, Ochalek A, Szczesna K, Avci HX, Kobolak J, et al. (2016) The positional identity of iPSC-derived neural progenitor cells along the anterior-posterior axis is controlled in a dosage-dependent manner by bFGF and EGF. Differentiation.

60. Zhu W, Gramlich OW, Laboissonniere L, Jain A, Sheffield VC, et al. (2016) Transplantation of iPSC-derived TM cells rescues glaucoma phenotypes in vivo. Proc Natl Acad Sci U S A 113: E3492-3500.

61. Chin MH, Mason MJ, Xie W, Volinia S, Singer M, et al. (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5: 111-123.

62. Ghosh Z, Wilson KD, Wu Y, Hu S, Quertermous T, et al. (2010) Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS One 5: e8975.

63. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, et al. (2010) Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. Cell Stem Cell 7: 249-257.

64. Lee JT (2012) Epigenetic regulation by long noncoding RNAs. Science 338: 1435-1439.

65. Dunham II, Kundaje AA, Aldred SFSF, Collins PJPJ, Davis CACA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

66. Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, FranCoijs K-J, et al. (2009) A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in Xenopus Embryos. Developmental Cell 17: 425-434.

67. Cotney J, Leng J, Oh S, Demare LE, Reilly SK, et al. (2012) Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. Genome Research 22: 1069-1080.

68. Xiao S, Xie D, Cao X, Yu P, Xing X, et al. (2012) Comparative epigenomic annotation of regulatory DNA. Cell 149: 1381-1392.

69. Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. Science 187: 226-232.

70. Riggs AD (1975) X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14: 9-25.

71. Riggs AD (2002) X chromosome inactivation, differentiation, and DNA methylation revisited, with a tribute to Susumu Ohno. Cytogenet Genome Res 99: 17-24.

72. Robertson KD (2001) DNA methylation, methyltransferases, and cancer. Oncogene 20: 3139-3155.

73. Baylin SB, Herman JG (2000) DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet 16: 168-174.

74. Mendelsohn AR, Larrick JW (2013) The DNA methylome as a biomarker for epigenetic instability and human aging. Rejuvenation Res 16: 74-77.

75. Irier HA, Jin P (2012) Dynamics of DNA methylation in aging and Alzheimer's disease. DNA Cell Biol 31 Suppl 1: S42-48.

76. Mastroeni D, Grover A, Delvaux E, Whiteside C, Coleman PD, et al. (2010) Epigenetic changes in Alzheimer's disease: decrements in DNA methylation. Neurobiol Aging 31: 2025-2037.

77. Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, et al. (2012) Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. Molecular Psychiatry 18: 340-346.

78. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-322.

79. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genetics 39: 457-466.

80. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480: 490-495.

81. Bird AP (1987) CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics.

82. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. Journal of molecular biology 196: 261-282.

83. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. (2012) The accessible chromatin landscape of the human genome. Nature 489: 75-82.

84. Drong AW, Nicholson G, Hedman AK, Meduri E, Grundberg E, et al. (2013) The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. PLoS One 8: e55923.

85. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, et al. (2010) Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet 86: 411-419.

86. Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, et al. (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464: 1082-1086.

87. Lindroth AM, Park YJ, McLean CM, Dokshin GA, Persson JM, et al. (2008) Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. PLoS Genet 4: e1000145.

88. Brinkman AB, Gu H, Bartels SJJ, Zhang Y, Matarese F, et al. (2012) Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Research 22: 1128-1138.

89. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, et al. (2011) Identification of genetic elements that autonomously determine DNA methylation states. Nature Genetics 43: 1091-1097.

90. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature: 1-5.

91. Hernando-Herraez I, Prado-Martinez J, Garg P, Fernandez-Callejo M, Heyn H, et al. (2013) Dynamics of DNA methylation in recent human and great ape evolution. PLoS Genetics 9: e1003763.

92. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, et al. (2013) Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. The American Journal of Human Genetics 93: 876-890.

93. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

94. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

95. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature: 1-9.

96. Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. Nature Reviews Genetics 10: 295-304.

97. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21: 447-455.

98. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, et al. (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genetics 8: e1002629.

99. Caliskan M, Cusanovich DA, Ober C, Gilad Y (2011) The effects of EBV transformation on gene expression levels and methylation profiles. Human Molecular Genetics 20: 1643-1652.

100. Grafodatskaya D, Choufani S, Ferreira JC, Butcher DT, Lou Y, et al. (2010) EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. Genomics 95: 73-83.

101. Tate PH, Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. Current Opinion in Genetics & Development 3: 226-231.

102. Hahn MA, Wu X, Li AX, Hahn T, Pfeifer GP (2011) Relationship between Gene Body DNA Methylation and Intragenic H3K9me3 and H3K36me3 Chromatin Marks. PLoS ONE 6: e18844.

103. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13: 484-492.

104. Tung J, Barreiro LB, Johnson ZP, Hansen KD, Michopoulos V, et al. (2012) Social environment is associated with gene regulatory variation in the rhesus macaque immune system. Proceedings of the National Academy of Sciences.

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

106. Yu M, Hon GC, Szulwach KE, Song C-X, Zhang L, et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149: 1368-1380.

107. Krueger F (2011) Bismark, A flexible aligner and methylation caller for Bi-sulfate seqencing applications. 1-3.

108. Guan Y, Stephens M (2008) Practical Issues in Imputation-Based Association Mapping. PLoS Genet 4: e1000279.

109. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100: 9440-9445.

110. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

111. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1-13.

112. Kim K, Doi A, Wen B, Ng K, Zhao R, et al. (2010) Epigenetic memory in induced pluripotent stem cells. Nature 467: 285-290.

113. Bar-Nur O, Russ HA, Efrat S, Benvenisty N (2011) Epigenetic memory and preferential lineage-specific differentiation in induced pluripotent stem cells derived from human pancreatic islet beta cells. Cell Stem Cell 9: 17-23.

114. Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, et al. (2010) Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol 28: 848-855.

115. Ohi Y, Qin H, Hong C, Blouin L, Polo JM, et al. (2011) Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nat Cell Biol 13: 541-549.

116. Kim K, Zhao R, Doi A, Ng K, Unternaehrer J, et al. (2011) Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol 29: 1117-1119.

117. Cahan P, Daley GQ (2013) Origins and implications of pluripotent stem cell variability and heterogeneity. Nat Rev Mol Cell Biol 14: 357-368.

118. Ma H, Morey R, O'Neil RC, Yupeng H, Brittany D, et al. (2014) Abnormalities in human pluripotent cells due to reprogramming mechanisms. Nature.

119. Kvaratskhelia M, Sharma A, Larue RC, Serrao E, Engelman A (2014) Molecular mechanisms of retroviral integration site selection. Nucleic Acids Research.

120. LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, et al. (2014) MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Research.

121. Okita K, Matsumura Y, Sato Y, Okada A, Morizane A, et al. (2011) A more efficient method to generate integration-free human iPS cells. Nat Meth 8: 409-412.

122. Cheng L, Hansen NF, Zhao L, Du Y, Zou C, et al. (2012) Low incidence of DNA sequence variation in human induced pluripotent stem cells

generated by nonintegrating plasmid expression. Cell Stem Cell 10: 337-344.

123. Rouhani F, Kumasaka N, de Brito MC, Bradley A, Vallier L, et al. (2014) Genetic Background Drives Transcriptional Variation in Human Induced Pluripotent Stem Cells. PLoS Genet 10: e1004432.

124. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, et al. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 30: 1363-1369.

125. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biology 13: R44.

126. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.

127. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Banovich NE, Kagan CL, et al. (2014) Generation of a Panel of Induced Pluripotent Stem Cells From Chimpanzees: a Resource for Comparative Functional Genomics.

128. Chen G, Gulbranson DR, Hou Z, Bolin JM, Ruotti V, et al. (2011) Chemically defined conditions for human iPSC derivation and culture. Nat Meth 8: 424-429.

129. Howden SE, Wardan H, Voullaire L, McLenachan S, Williamson R, et al. (2006) Chromatin-binding regions of EBNA1 protein facilitate the enhanced transfection of Epstein-Barr virus-based vectors. Hum Gene Ther 17: 833-844.

130. Müller F-J, Schuldt BM, Williams R, Mason D, Altun G, et al. (2011) A bioinformatic assay for pluripotency in human cells. Nature Methods 8: 315-317.

131. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571-1572.

132. Du P, Kibbe WA, Lin SM (2008) lumi: a pipeline for processing Illumina microarray. Bioinformatics 24: 1547-1548.

133. WJ K, CW S, TS F, KM R, TH P, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

134. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011: bar049.

135. Köster J, Rahmann S (2012) Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28: 2520-2522.

136. Kim K-Y, Hysolli E, Tanaka Y, Wang B, Jung Y-W, et al. (2014) X Chromosome of Female Cells Shows Dynamic Changes in Status during Human Somatic Cell Reprogramming. Stem Cell Reports 2: 896-909.

137. Ernst J, Pouya K, Tarjei SM, Noam S, Lucas DW, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43-49.

138. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PAC, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.

139. Hochedlinger K, Jaenisch R (2015) Induced Pluripotency and Epigenetic Reprogramming. Cold Spring Harb Perspect Biol 7.

140. Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Gallego Romero I, et al. (2016) Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory Differences in iPSCs. PLoS Genet 12: e1005793.

141. Kilpinen H, Goncalves A, Leha A, Afzal V, Ashford S, et al. (2016) Common genetic variation drives molecular heterogeneity in human iPSCs. bioRxiv.

142. Thomas SM, Kagan C, Pavlovic BJ, Burnett J, Patterson K, et al. (2015) Reprogramming LCLs to iPSCs Results in Recovery of Donor-Specific Gene Expression Signature. PLoS Genet 11: e1005216.

143. van de Geijn B, McVicker G, Gilad Y, Pritchard JK (2015) WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods 12: 1061-1063.

144. Flutre T, Wen X, Pritchard J, Stephens M (2013) A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet 9: e1003486.

145. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, et al. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 22: 1798-1812.

146. Spandidos DA, Anderson ML (1984) A tissue-specific transcription enhancer element in the human immunoglobulin lambda light chain locus. FEBS Lett 175: 152-158.

147. Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 12: 283-293.

148. Nakabayashi H, Koyama Y, Suzuki H, Li HM, Sakai M, et al. (2004) Functional mapping of tissue-specific elements of the human alpha-fetoprotein gene enhancer. Biochem Biophys Res Commun 318: 773-785.

149. Jongens TA, Fowler T, Shermoen AW, Beckendorf SK (1988) Functional redundancy in the tissue-specific enhancer of the Drosophila Sgs-4 gene. EMBO J 7: 2559-2567.

150. Gillies SD, Morrison SL, Oi VT, Tonegawa S (1983) A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell 33: 717-728.

151. Mali P, Yang L, Esvelt KM, Aach J, Guell M, et al. (2013) RNA-guided human genome engineering via Cas9. Science 339: 823-826.

152. Lian X, Zhang J, Azarin SM, Zhu K, Hazeltine LB, et al. (2013) Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. Nat Protoc 8: 162-175.

153. Burridge PW, Matsa E, Shukla P, Lin ZC, Churko JM, et al. (2014) Chemically defined generation of human cardiomyocytes. Nat Methods 11: 855-860.

154. Marroquin LD, Hynes J, Dykens JA, Jamieson JD, Will Y (2007) Circumventing the Crabtree effect: replacing media glucose with galactose increases susceptibility of HepG2 cells to mitochondrial toxicants. Toxicol Sci 97: 539-547.

155. Chan YC, Ting S, Lee YK, Ng KM, Zhang J, et al. (2013) Electrical stimulation promotes maturation of cardiomyocytes derived from human embryonic stem cells. J Cardiovasc Transl Res 6: 989-999.

156. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10: 1213-1218.

157. Ongen H, Buil A, Brown A, Dermitzakis E, Delaneau O (2015) Fast and efficient QTL mapper for thousands of molecular phenotypes. bioRxiv.

158. Alexa A, J R (2016) topGO: Enrichment Analysis for Gene Ontology. R package version 2240.

159. Kyttala A, Moraghebi R, Valensisi C, Kettunen J, Andrus C, et al. (2016) Genetic Variability Overrides the Impact of Parental Cell Type and Determines iPSC Differentiation Potential. Stem Cell Reports 6: 200-212.

160. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, et al. (2015) Genomic variation. Impact of regulatory variation from RNA to protein. Science 347: 664-667.