THE UNIVERSITY OF CHICAGO

INDUCED PLURIPOTENT STEM CELLS AS A MODEL TO STUDY INDIVIDUAL VARIATION AND COMPARATIVE GENOMICS

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

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: HUMAN GENETICS

BY

SAMANTHA MARIE THOMAS

CHICAGO, ILLINOIS

DECEMBER 2016

TABLE OF CONTENTS

ABSTRACT

The past decade of genetics research has been defined by the discovery of the profound effects non-coding genetic variation can have on the phenotypes that distinguish humans from each other and from our close evolutionary relatives. The full implications of this new understanding are largely unexplored, however, as modern ethics restricts experimentation in humans and most primates, rendering data from dynamic processes almost non-existent. The study of regulatory molecular dynamics has been changed entirely by the availability of protocols to generate iPSCs and differentiate them into adult cell types. The molecular basis of disease mechanisms, drug response, and developmental processes can now be studied in the relevant tissue, presenting an overwhelming spectrum of possible applications. Of particular interest to comparative biologists, long-standing questions about the relative conservation of early developmental states can now, for the first time, be ethically explored in closely related primates. In this dissertation, we first discuss evidence that iPSCs can faithfully model genetic variation, even when sourced from highly dysregulated cells. We then use an iPSC-based model to study the temporal profile of conservation between humans and chimpanzees during early endoderm development and identify patterns of divergence over developmental stages.

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible if it were not for the incredibly supportive environment it was conducted in. I am deeply indebted to members of the Gilad lab who have been like a scientific family to me for the past four years. Cell culture inherently comes with many difficulties and the camaraderie of Nicholas Banovich, Courtney Kagan, Bryan Pavlovic, and Michelle Ward, the iPSC group, was invaluable. I am particularly grateful to Bryan Pavlovic for countless hours of troubleshooting and optimizing to bring a project to life that at times seemed impossible. I would also specifically like to thank John Blischak for many discussions about the difficult statistical issues that we all dread confronting. John's clear thinking and patience were a great help to me. Likewise, I am appreciative for Lauren Blake's time thinking through every possible way to normalize and model time-course data with me. Most of all, I would like to sincerely thank my advisor Dr. Yoav Gilad. Dr. Gilad has been unwavering in his support and I have learned so much from his optimism and fearless approach to science. Aside from making the lab a lively and engaging place to work, his attitude also encouraged a culture of curiosity and bravery of scientific thought, which I particularly benefited from. Working in comparative genomics was one of the most interesting things I have ever had a chance to do, and I would have never tried had I been mentored by anyone else. Dr. Gilad was welcoming to his field of primate research and through many hours of discussing what was most interesting and important in our work, and teaching me by example about narrative style, he helped me develop a voice in science. For this especially, I could not be more grateful. Looking forward, I

can already see that I will miss the vibrant spirit of the lab and this remarkable group of scientists.

Many others have made the years of PhD work constructive, fun, and memorable. The graduate program administrators Sue Levison, Elise Covic, Alison Anastasio, Shay McAllister, Sarah Blum, and Candice Lewis have created the best support system I could have asked for. Likewise, I am indebted to Dr. Jim Woodruff at Pritzker and Emily Ho at UCSD. I also greatly benefited from the direction of my committee members Dr. Marcus Clark, Dr. Eileen Dolan, Dr. Matthew Stephens, and Dr. Minoli Perera whose thoughtful advice over the years has shaped this thesis. I am lucky to have worked in the Human Genetics department and to have had the opportunity to collaborate with other current and former HG professors Dr. Vinny Lynch and Dr. Jonathan Pritchard. I am also grateful for such a close MSTP class in Anya Bershad, Jennifer Jacobsen, and Alan Hutchison. Lastly, I am so fortunate to have a wonderful family, Mary, David, and Ben, who have always supported me.

CHAPTER ONE

Introduction

An unexpected finding of the genome-wide association study research over the past decade has been the predominance of non-coding mutations in trait associated variants[1]. This result is mirrored by the discovery in comparative biology that a significant fraction of DNA sequence under selection in the human lineage does not code for protein[2]. Together, these findings have led to the general understanding that many essential human traits arise from alterations in gene regulation, not from sequence differences in genes themselves. One of the most important tasks for biologists now is to understand how regulatory variation gives rise to phenotypes of interest.

Addressing these questions has required a shift in technology away from single protein assays to methods that read out the entire landscape of the genome, epigenome, and transcriptome. Over the past decade, approaches have been developed to map the genomic distribution of epigenetic marks, locate occupied transcription factor binding sites, and modify DNA with high specificity. To keep pace with data sets that are growing in size and complexity, statistical tools have also matured and many methods for analyzing regulatory relationships in high-dimensional data are now freely available.

As data acquisition and analysis techniques have become more sophisticated, however, the absence of a satisfactory model system for studies of gene regulation has remained a serious limitation. Such a model must approximate the relevant physiological conditions closely enough

to capture the regulatory effects of non-coding variation, which may only be active in a particular tissue, developmental time-point, or physiological perturbation[3]. In an attempt to expose and investigate the appropriate cellular state, geneticists have made use of a number of models, whose advantages and limitations will be briefly reviewed, and now demonstrate an increasing preference for the iPSC system.

1.1 History of Models for Studies of Gene Regulation

Due to an early awareness of the tissue-specificity of non-coding regulation, the first gold standard material for gene expression studies was a direct tissue biopsy. This approach has produced many intriguing findings in comparative biology and medical research. Microarray studies in tissue from humans and other primates, for instance, have discovered the tissue specific nature of directional selection on gene expression levels[4] and highlighted control of transcription factor expression levels as a particular target of evolutionary pressures[5]. In medicine, gene expression studies in diseased tissue compared to healthy controls have been essential to understand progression of pathology[6-9], and are still used to stage certain diseases, particularly cancers[10; 11]. However, in an effort to work ethically with donors, most samples from humans and other primates are adult tissues obtained post-mortem, and often frozen. They are therefore not a good option for studying the dynamics of gene regulation, for instance during development or in response to physiological perturbations. In addition, donated tissues are influenced by the donor's lifestyle in unpredictable and potentially severely confounding ways. Thus, most conclusions from such studies are preliminary and clearly caveated.

Instead, human genetic disease research has historically relied heavily on animal models, whose environment can be controlled and manipulated. This method has been especially popular for monogenic diseases, which are theoretically possible to model in animals using a gene knockout approach or by identification of an allele that tracks with the disease phenotype. In fact, a number of disorders have been adequately reproduced in mice, including monogenic diabetes and even a polydactyly syndrome caused by regulatory mutations[12]. Despite these intermittent successes however, the human and murine lineage diverged around 100 million years ago and their genomes are likely to differ markedly in the pathways being investigated, a concern supported by the fact that knockout mice often do not manifest the expected clinical symptoms[13]. For example, mouse models of cystic fibrosis die of intestinal disease before they exhibit any obstructive airway phenotype[14]. In addition, it is unclear whether certain human diseases could ever be modeled in animals. This is particularly true for complex disorders, such as Alzheimer's or heart disease, that arise from the constellation of many genetic and environmental risk factors. In these situations, samples from human individuals harboring the condition are invaluable.

The limitations of tissue and animal work are well recognized and have motivated the widespread use of immortalized human cell lines, such as lymphoblastoid cell lines (LCLs), for genetic studies. Because LCLs are a renewable cell type, they are a good model for experiments and have been essential in developing our current understanding of human gene regulation[15-20]. LCLs are easy to establish from blood and are available in large panels from diverse populations including the HapMap participants and cohorts of patients with genetic disorders[10; 21-24]. However, LCLs are an artificial system, not a naturally occurring cell type, and the

immortalization process induces widespread epigenetic changes, obscuring the innate gene regulation[25; 26]. Moreover, at best LCLs can only be expected to represent one human cell type, the B-cells from which they were derived. While LCLs are a convenient model, these are serious limitations and for more sensitive investigations, another resource is needed.

1.2 Development of the iPSC Model

Over the past decade, pluripotent stem cells (PSCs) have emerged as a viable system for studying dynamic gene regulation in a variety of human tissues. The term pluripotent stem cell includes any cell that is capable of giving rise to all three germ layers, endoderm, ectoderm, and mesoderm, as well as the hundreds of cell types arising each lineage. Because they can be expanded and differentiated in live culture, PSCs are a great option for experiments previously only practical in immortalized cell lines. PSCs have a clear advantage over transformed cell lines, though, because of their theoretical ability to give rise to any cell type in the adult body, granting access to a larger scope of biology.

Although pluripotent embryonic stem cells have been used by a few groups since the late 1990s, PSCs only reached widespread acceptance in 2007 when Yamanaka et al. developed a method to induce pluripotency in adult human somatic cells by transfecting them with four pluripotency related transcription factors[27]. This advance shifted the source of PSCs from the limited, ethically-charged resource of embryonic stem cells, to any adult willing to donate a blood or skin sample. The resulting induced pluripotent stem cells (iPSCs) were quickly recognized for their potential to transform drug discovery, regenerative medicine, and

personalized medicine. Furthermore, because iPSCs retain the genotype of their donor, they are also a powerful resource for basic human genetics research.

Two fields of human genetics particularly welcoming of the iPSC model are genetic disease research and comparative biology. Both disciplines have long suffered from the lack of environmentally controlled tissue samples in which to study the regulatory mechanisms of heritable traits. Those researching genetic diseases can now efficiently derive iPSCs directly from patient blood or from the many banks of LCLs that have been established for a given disease. In fact, an international movement has recently announced its intention to coordinate the generation of human iPSC panels from thousands of patients with a disease of interest and healthy controls[28]. For comparative studies, panels of iPSCs have now been created from several mammalian species including our closest evolutionary relative the chimpanzee[29]. Though the full potential of iPSCs to shape these fields is years away, impressive strides have already been made in both disciplines, and are summarized here.

1.3 iPSCs to study the mechanisms of genetic disease

Perhaps the most highly anticipated feature of iPSCs is their theoretical ability to give rise to clinically important cell types that are prohibitively difficult to obtain and experiment on, such as cardiac tissue, pancreatic islet cells, and neurons. Although many disorders arising from the function of these cells are understood to have a genetic basis, regulatory mechanisms underlying the genetic predisposition are almost entirely unstudied in these cell populations. Thus, as protocols are being developed to produce these cell types, disease modeling and functional genetic studies are being performed just as quickly.

iPSCs have now been generated from hundreds of patients with monogenic diseases including those with Huntingtons[30; 31], Marfan's syndrome[32], and familial Alzheiemer's[33; 34], all serious, incurable, poorly understood disorders. Many of these studies have been successful in reproducing known clinical pathology, such as the accumulation of the Aβ oligomer in iPSC-derived neurons made from an Alzheimer's patient, and have provided further resolution of the molecular mechanisms that lead to these phenotypes. One recent study exemplified the promise of iPSCs by using zinc-finger nuclease mediated gene targeting to correct a genetic defect thought to cause familial ALS development[35]. Through subsequent dissection of the transcriptional and functional consequences of the sequence correction, the authors identified a novel cellular pathway whose dysregulation in motor neurons may be the key process underlying this fatal neurological disease.

The iPSC system is also uniquely suited to the study of complex genetic diseases, arising from the interaction of multiple genetic and environmental factors. Recently iPSC derived neurons were generated from patients with schizophrenia, a complex psychiatric disorder with a high heritability but an unknown genetic origin. Gene expression studies in these neurons were the first to be performed in a controlled experimental setting, as opposed to post-mortem tissue, and flagged new pathways not yet identified by GWAS hits, along with several known ones[36]. Such approaches are also being undertaken for autism[37], sporadic Alzheimer's[38], and diabetes[39] with the ultimate goal of finally resolving the mechanisms behind their respective genetic risk factors. The concept of using iPSCs to probe genetic susceptibility to adverse drug reactions is another popular idea recently realized by several carefully designed studies in the relevant iPSC-derived tissue. For instance, the effect of a genetic variant in the gene *VAC14* on sensitivity

to chemotherapeutic agents docetaxel and paclitaxel was assessed in knockdown iPSC-derived peripheral and cortical neurons, validating a GWAS hit for susceptibility to drug-induced neuropathy[40; 41]. iPSC-derived cardiomyocytes have also been demonstrated to successfully recapitulate predisposition to doxorubicin-induced cardiotoxicity, indicating they may be a good option for studying the genetic basis of susceptibility, [37]

1.4 iPSCs to study the regulatory mechanisms of human evolution

Overwhelming evidence suggests that noncoding sequence changes contribute significantly to phenotypic differences between humans and other primates[2; 42-48]. Noncoding elements are thought to be particularly responsive to evolutionary pressures because their effects are often specific to a single tissue and therefore less likely to be deleterious[43]. For this same reason, however, the consequences of noncoding differences between species have been far more challenging to study than those of coding mutations, which do not require access to the relevant tissue and can generally be inferred from the sequence itself. Such sequence based methods have been optimistically employed to predict the functional significance of noncoding variation, however the process has been computationally intensive and has not resulted in the same degree of confidence in its predictions.

Comparative biologists, especially those studying primates, have turned their attention to iPSC-derived cells to study tissue-specific effects of noncoding sequence changes across lineages[49; 50]. In an excellent illustration of this approach, Prescott et al. recently reported divergent use of regulatory elements in human and chimp iPSC-derived neural crest cells, tracing genome-wide epigenetic differences between the species to a minimal set of sequence variants in

transcription factor binding sites near genes known to be active in facial morphology[51]. Another intriguing comparative study used a panel of iPSCs and embryonic stem cells (ESCs) to model differences in early cortical development between primates[52], a phenotype whose genetic basis has long been a topic of intense interest and undoubtedly soon will be explored in this iPSC model.

As highlighted by these recent applications, one of the most important advantages of using iPSCs to study speciation is access to early developmental time points. It is thought that many of the most profound differences between species are established early in embryogenesis, though during these initial stages no morphological differences can be appreciated[50]. Decades of research at the intersection of evolution and development (evo-devo) have revealed that differential spatio-temporal regulation of conserved genes during development underlies even the most extreme phenotypic differences between animals[53]. The large body of evo-devo research of this nature in drosophila can now be complimented by developmental studies in iPSCs from other animals. In chapter two, we use iPSCs derived from humans and chimpanzees to compare regulatory processes involved in the endoderm development.

1.5 Limitations of iPSC model for studies of gene regulation

Although iPSCs are a promising new resource, they are currently limited in several ways. The first consideration which has received significant attention is the possibility of residual epigenetic signatures of the cells from which the iPSCS were derived, often LCLs or fibroblasts. The extent to which this artifact occurs is debated, with some studies observing a significant lingering effect of the original cell type[54; 55]. However our group has carefully studied the

methylation and expression patterns in iPSCs generated from multiple cell types provided by the same individual and concluded that when properly controlling for background genotype, the effect on global studies of gene regulation is likely to be very small[56].

Another concern is whether the disruption of gene regulation inherent to cultured or immortalized cell lines, a large source of starting material for iPSCs, persists through the reprogramming process. Presumably the most seriously dysregulated starting cell type for iPSCs are LCLs, immortalized by transformation with Epstein-Barr virus, and well-known to acquire regulatory artifacts as the cell lines adapt to prolonged cell culture. In chapter one, we report that iPSCs better reflect the donor gene expression signature than the LCLs from which they were derived suggesting that reprogramming transformed cells restores some of the innate genetic control of gene regulation.

Perhaps the most significant, albeit temporary, limitation of iPSCs is the early state of most tissue differentiation protocols. Although iPSCs are theoretically capable of giving rise to any adult cell type, no protocol has managed to achieve perfect recapitulation of mature human tissue. This limits the conclusions of current studies of gene regulation in iPSC-derived tissue to a fetal or adult-like context. This is of particular concern for studies of mechanisms underlying late-onset diseases. As differentiation protocols become more advanced, however, the impact of this issue will diminish.

1.6 Conclusion

The rate at which iPSC differentiation protocols and assays to functionally study gene regulation are being generated reflects a deep excitement for their potential to dramatically improve our

understanding of human traits. This thesis provides a report on the suitability of LCL-derived iPSCs for this application and a framework for using iPSCs to study comparative development.

CHAPTER 2

Reprogramming LCLs to iPSCs Results in Recovery of Donor-Specific Gene Expression Signature

2.1 Introduction

Renewable cell models are widely recognized as valuable platforms for studies of human genotype-phenotype interactions because they are easily manipulated, scalable, and are specific to human physiology (in contrast to lab animal models). Epstein-Barr virus (EBV) transformed lymphoblastoid cell lines (LCLs) are one such commonly-used model. In recent years, LCLs have been used to study genetic influence on disease traits [21], drug response [41; 57-64], and gene regulation [15; 19]. In particular, much of what we now know about associations of human genetic variation with differences in gene regulation is based on studies that used data from LCLs. There is little doubt that many fundamental regulatory principles that we have learned by generating and analyzing data from LCLs are generally shared with primary tissues. However, a critical property of any *in vitro* cellular model is the ability to faithfully recapitulate the specific regulatory properties of the donor's primary tissue. In that regard, though LCLs have clearly been a convenient and useful model, there is concern that factors related to immortalization and cell line maintenance obscure genetic signal in LCLs [65-67].

A number of studies have characterized differences in gene regulatory phenotypes between LCLs and primary tissues [26; 68-71]. These have shown that a large number of genes are differentially expressed between primary cells and cell lines, and that thousands of CpG sites

are differentially methylated between LCLs and primary blood cells. Our group has also demonstrated disruptions in gene regulation in LCLs by studying multiple independent replicates of LCLs from isolated primary B cells of six individuals and repeatedly subjecting the cell lines to cycles of freeze, thaw, and recovery. We found that newly transformed LCLs (within a few passages after the EBV transformation) largely maintained individual differences in gene expression levels. However, LCLs that had been frozen and thawed at least once (we referred to these as mature LCLs) exhibited a substantial loss of inter-individual variation in gene expression levels [25; 26].

On the one hand, it is unlikely that the loss of the donor effect on gene expression would lead to false positive findings of genetic influence on gene regulation. Indeed, we reported that genes associated with previously identified eQTLs retain relatively high variation in gene expression levels between individuals even after repeated freeze-thaw culturing cycles. Yet on the other hand, because much of the individual variation observed in primary tissues is not exhibited by LCLs, studies using the LCL model are limited in their ability to detect donor differences.

The induced pluripotent stem cell (iPSC) system is another renewable cell model that is increasingly used to study individual phenotypic variation because it can ultimately provide access to a wide range of tissue types through the use of differentiation protocols. However, the capacity of iPSCs and derived cell types to faithfully recapitulate *in vivo* physiology is also still largely unknown. Previous studies have noted a significant effect of donor on traits in iPSCs such as hematopoietic [72], neuronal [73] and hepatic [74] differentiation potential. Importantly, the genetic background of iPSCs generated from peripheral blood mononuclear cells and fibroblasts

was recently demonstrated to account for more of the variation in gene expression between iPSC lines than any other tested factor such as cell type of origin or reprogramming method [75]. While these findings indicate that reprogramming iPSCs from primary tissues preserves individual variation in gene expression, it is unknown whether reprogramming highly manipulated immortalized cell lines, such as LCLs, to iPSCs can recover the individual gene expression patterns lost during cell line maintenance.

Because LCLs are available in large banks representing disease populations or ethnicities, they are a promising source of starting material for iPSC generation if disruptions in gene regulation do not persist through the reprogramming process. In the present study, we ask whether reprogramming mature LCLs to iPSCs can result in the recovery of individual variation in gene expression that had been lost during the LCL maturation and maintenance process.

2.2 Results

To test whether reprogramming LCLs to iPSCs could recover the effect of donor on gene expression profiles, we generated iPSCs from three mature LCLs of each of six Caucasian individuals for a total of 17 pairs of cell lines (one iPSC line failed to reach the requisite ten passages and was excluded from the study; see methods). We have previously collected gene expression data from the LCLs at earlier stages [25; 26]. For the current study, we quantified whole genome gene expression microarray data from the 17 mature LCLs immediately prior to reprogramming and from stable and validated iPSCs. See Fig. 1 for schematic of the study design and Table 1 for the processed gene expression data from all samples.

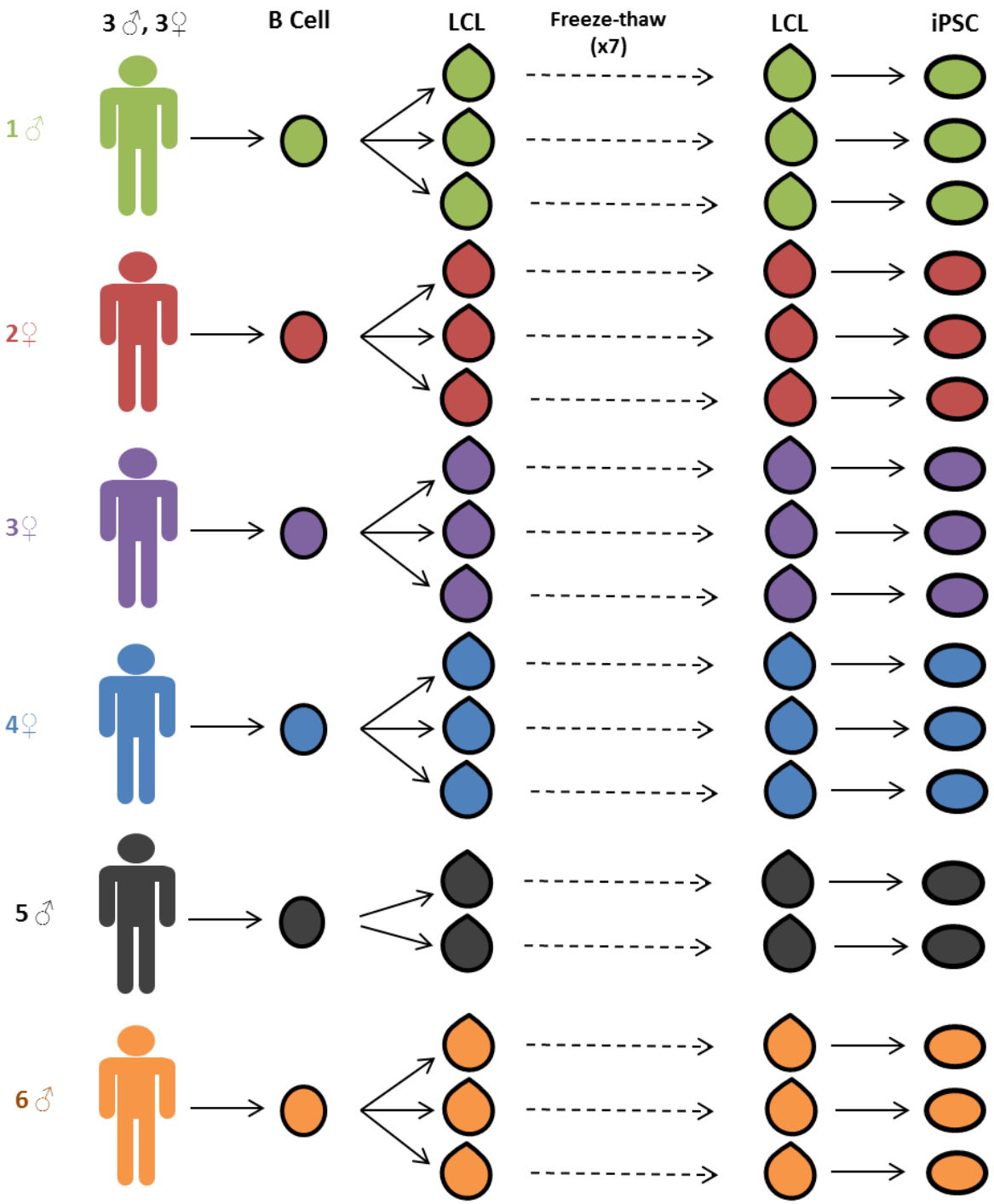
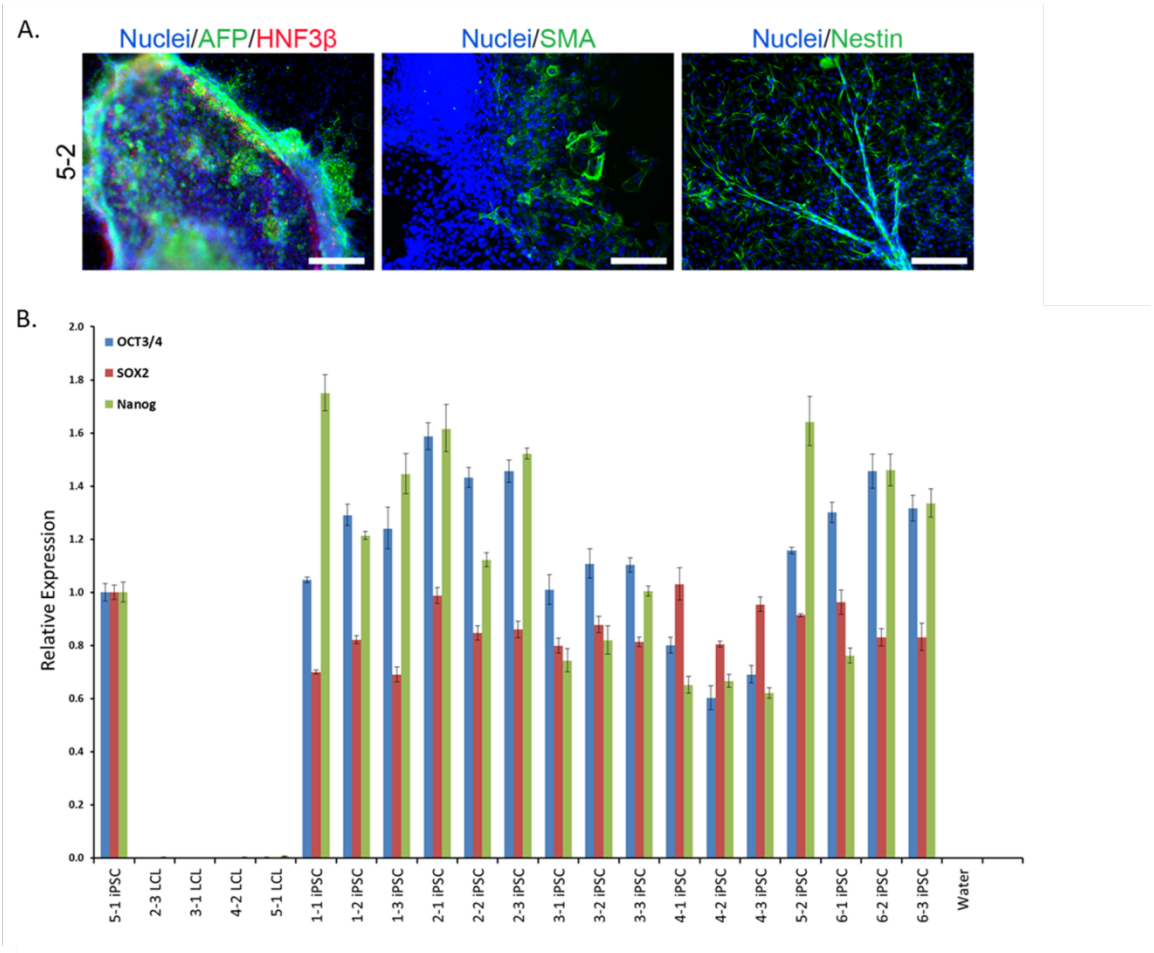


Figure 2.1: Study design**.** Three independent lymphoblastoid cell lines (LCLs) were generated for each of six unrelated Caucasian individuals. LCLs were frozen and thawed seven times. After the seventh thaw, the LCLs were reprogrammed to iPSCs. Gene expression data was collected from LCLs immediately before reprogramming and from stable iPSC lines.

Generation and Validation of the iPSCs

We reprogrammed mature LCLs, which had previously undergone seven freeze-thaw culturing cycles, to iPSCs using an episomal transfection approach [76-78] (see Methods for more details). We reprogrammed the LCLs in four batches; scheduling LCLs derived from the same individual to different reprograming batches to ensure that no artificial correlation structure was introduced between 'reprograming batch' and 'donor individual' in the process of iPSC generation. All iPSC lines were confirmed to be pluripotent using an embryoid body assay (Fig. 2A,C.), qPCR for pluripotency-associated transcription factors (Fig. 2B), genomic PCR to confirm the absence of reprogramming plasmids (Fig. 3 & 4), and PluriTest, a bioinformatic classifier designed to assess pluripotency using gene expression data [79] (Table 2). Three independently established LCLs were successfully reprogrammed into validated iPSCs for all but one individual, for which only two iPSC lines were obtained.

Figure 2.2: iPSC generation and validation.



C.

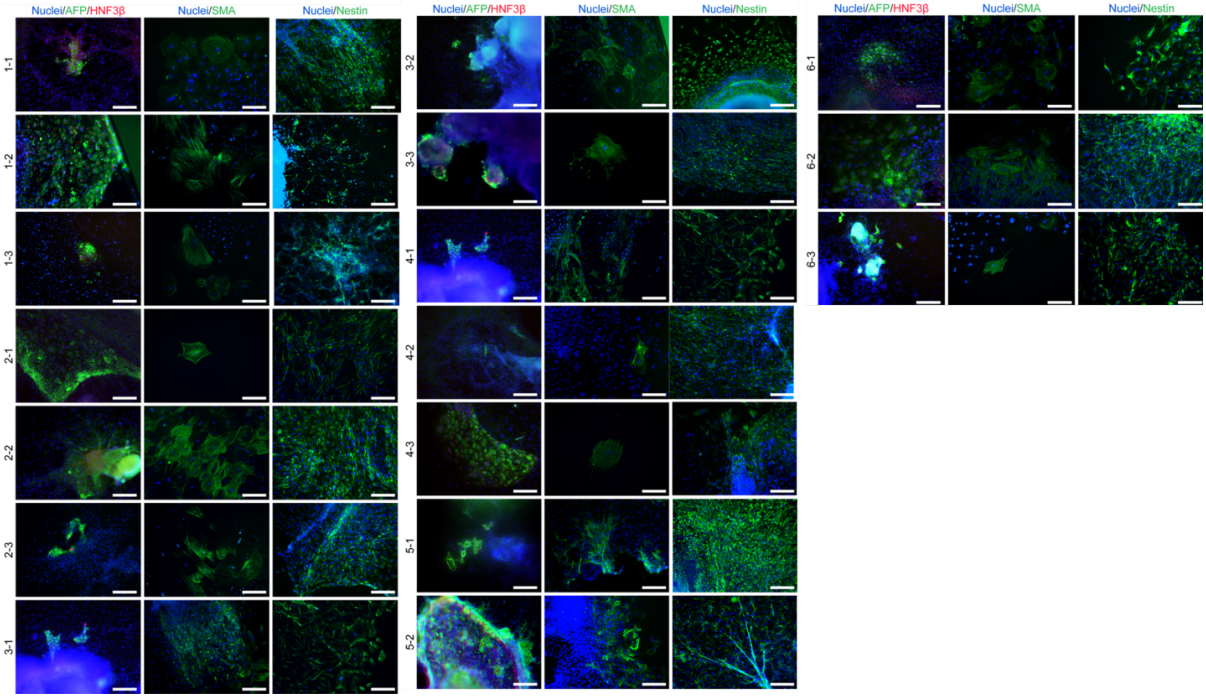


Figure 2.2 continued: iPSC generation and validation. A. Representative embryoid body staining for iPSC line 5-2 demonstrating differentiation potential for endoderm, mesoderm, and ectoderm lineages. Scale bars represent 200 µm B. Results from qPCR for three endogenous pluripotency-related transcription factors, normalized to GAPDH. iPSC 5-1 was randomly chosen as a reference sample. C. Embryoid body assay: Immunocytochemistry approach to test for a cell line's ability to spontaneously differentiate through endoderm: HNF3β and α-fetoprotein (AFP), mesoderm: smooth muscle actin (SMA), and ectoderm: nestin lineages. Scale bar: 200 µm. Individual channel levels, brightness, and contrast were adjusted using Adobe Photoshop CS6

Name	Forward	Reverse	Use	Source of Primer:
GAPDH	ACCAACTGCTTAGCACCCCTGG	ATCACGCCACAGTTTCCCGGAG	Normalizing gene	Romero et. al. http://dx.doi.org/10.1101/008862
OCT3/4(CDS)	CCCCAGGGCCCCATTTTGGTACC	ACCTCAGTTTGAATGCATGGGAGAGC	Endogenous qPCR	PMID: 21460823
SOX2 (CDS)	TTCACATGTCCCAGCACTACCAGA	TCACATGTGTGAGAGGGGCAGTGTGC	Endogenous qPCR	PMID: 21460823
NANOG_CDR	CAGAAGGCCTCAGCACCTAC	ATTGTTCCAGGTCTGGTTGC	Endogenous qPCR	PMID:18029452
EBNA1_End1	ATC ATC ATC CGG GTC TCC ACC G	ATT GCA GGT AGG AGC GGG CTT TG	Genomic: Plasmid and EBV integration	B. Pavlovic
EBV_Genome_F	ATC AGG CTG CAT GTG CTC TTC CC	TTT GCA TGG TCA GGC TCC GTG TC	Genomic: EBV integration	B. Pavlovic
PCXLE_Common_F	TTTGCAAGCAGCAGATTACGCGCAG	TTTGCTCACCCAGAAACGCTGGTGAAAG	Genomic: Plasmid	B. Pavlovic

Table 2.1: PCR primer information: Sequence, use, and source of primers used for PCR and qPCR.

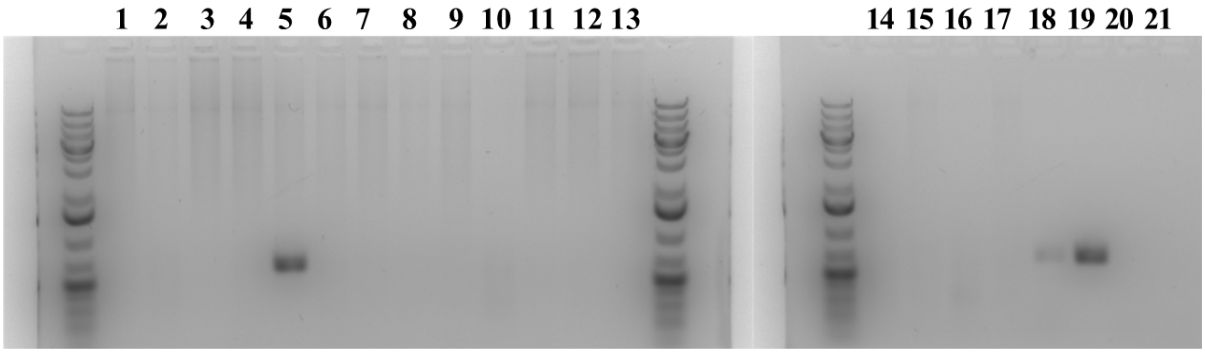


Figure 2.3: PCR for EBNA-1**:** Reprogramming vectors and Epstein-Barr virus are absent in all iPSC lines except 3-2 (see Figure S3)

1.4-1 iPSC, **2.** 6-3 iPSC, **3.** 5-2 iPSC, **4.** 3-3 iPSC, **5.** 3-2 iPSC, **6.** 6-2 iPSC, **7.** 2-3 iPSC, **8.** 5-1 iPSC, **9.** 1-2 iPSC, **10.** 1-1 iPSC, **11.** 3-1 iPSC, **12.** 3-3 iPSC, **13.** 2-2 iPSC, **14.** 2-1 iPSC, **15.** 4-2 iPSC, **16.** 1-3 iPSC, **17.** 6-1 iPSC, **18.** Reprogramming plasmids (positive control) **19.** LCL DNA (YRI lines 18508 and 19238, positive control). **20.** Fibroblast DNA (negative control), **21.** Water

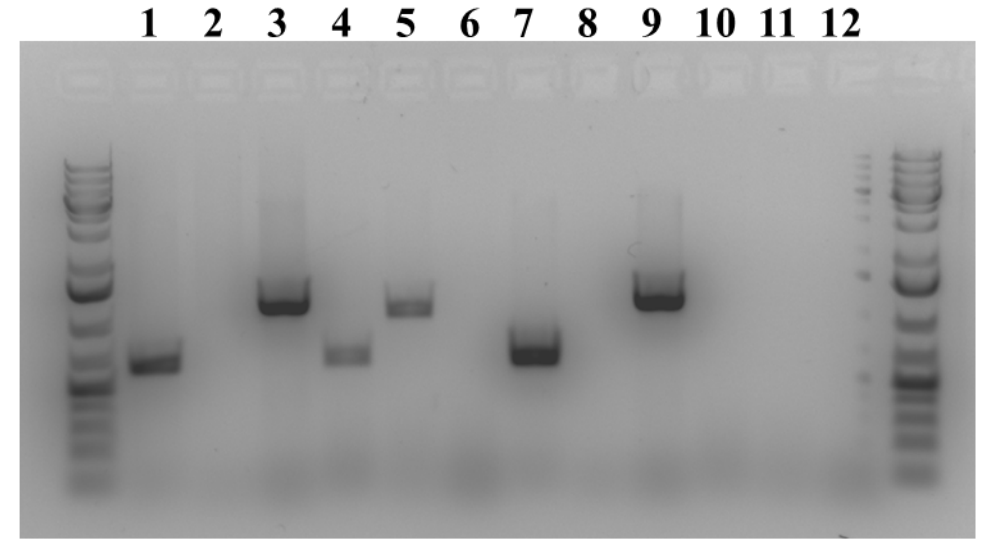


Figure 2.4: PCR for PCXLE (plasmid) and EBV genome for iPSC line 3-2: iPSC 3-2 exhibits presence of EBV and absence of reprogramming plasmids.

1.3-2 iPSC/EBNA-1 primer set, **2.** 3-2 iPSC/PXCLE primer set, **3.** 3-2 iPSC/EBV primer set, **4.** Reprogramming plasmid template/EBNA-1 primer set, **5.**Reprogramming plasmid Figure 2.4 continued: template/PXCLE primer set, **6.** Reprogramming plasmid template/EBV primer set, **7.** LCL DNA/EBNA-1 primer set, **8.** LCL DNA/PCXLE primer set, **9.** LCL DNA/EBV primer set. **10.** Fibroblast DNA/EBNA-1 primer set **11.** Fibroblast DNA/PCXLE primer set **12.** Fibroblast DNA/EBV primer set.

Recovery of the Individual Signature of Gene Regulation

We collected high quality RNA (RIN score range: 7.6-9.9; Table 2) from LCLs immediately prior to reprogramming and from the stable and validated iPSC lines after at least 10 passages (see Table 2 for specific passage information). We quantified gene expression levels for all samples using the Illumina Human HT12v4 microarray platform. As a first step of our analysis, we excluded data from probes whose target transcripts did not map to a unique Ensembl gene ID, those that spanned an exon-exon junction, and those that were not detected as 'expressed' in at least two samples from either cell type (we note that our general observations are robust with respect to a wide range of this inclusion criteria). We also excluded from the analysis data from probes with a known SNP with a minor allele frequency > 0.05 in the European population, based on the 1000 Genomes phase I data, to eliminate the possibility of an artificial effect of genotype on the hybridization-based estimates of gene expression levels. We then quantile-normalized the combined data from the remaining probes across all samples. We examined and corrected for array batch using the approach of Johnson et al [80] (see Methods). Finally, we obtained normalized expression levels for 12,243 genes detected as expressed in our samples (S1 Table). Using a linear model-based Empirical Bayes method (implemented in the

'limma' R package [81]), we classified 8,185 genes as differentially expressed between iPSCs and LCLs (FDR < 1%; see Methods for more details about modeling and hypothesis testing).

Name	Individual	Gender	Reprogrammed	Reprograming batch	Passage RNA collected	RNA collected	Extr Batch	RNA extracted	RIN	Array Batch	RMSD	Raw PluriScore
1-1 LCL	1	Male		0		8/19/2013	3	5/5/2014	9.9	1	0.98	-83.853
1-1 iPSC	1	Male	8/19/2013	1	13	12/10/2013	3	5/5/2014	8.8	1	0.421	33.074
2-1 LCL	2	Female		0		8/19/2013	3	5/5/2014	9.7	1	1.01	-78.286
2-1 iPSC	2	Female	8/19/2013	1	13	12/10/2013	3	5/5/2014	9.2	1	0.432	33.024
3-1 LCL	3	Female		0		8/19/2013	3	5/5/2014	9.9	1	0.99	-82.105
4-1 iPSC	4	Female	8/19/2013	1	13	12/10/2013	3	5/5/2014	9.3	1	0.418	34.24
4-1 LCL	4	Female		0		8/19/2013	3	5/5/2014	9.7	1	0.999	-79.884
3-1 iPSC	3	Female	8/19/2013	1	13	12/10/2013	3	5/5/2014	9.3	1	0.412	35.464
6-1 LCL	6	Male		0		1/13/2014	3	5/5/2014	9.7	1	0.996	-78.403
6-1 iPSC	6	Male	1/13/2014	4	11	4/8/2014	3	5/5/2014	9.2	1	0.399	31.304
1-2 LCL	1	Male		0		8/26/2013	1	5/2/2014	9.7	2	1.019	-78.863
1-2 iPSC	1	Male	8/26/2013	2	12	12/10/2013	1	5/2/2014	9.2	2	0.373	37.406
2-2 LCL	2	Female		0		8/26/2013	1	5/2/2014	9.6	2	1.017	-78.006
2-2 iPSC	2	Female	8/26/2013	2	12	12/10/2013	1	5/2/2014	9.5	2	0.381	34.422
3-2 LCL	3	Female		0		8/26/2013	1	5/2/2014	9.8	2	0.991	-78.423
3-2 iPSC	3	Female	8/26/2013	2	12	12/10/2013	1	5/2/2014	9.6	2	0.386	32.253
4-2 LCL	4	Female		0		9/4/2013	1	5/2/2014	9.8	2	0.994	-84.002
4-2 iPSC	4	Female	9/2/2013	3	11	12/10/2013	1	5/2/2014	9.3	2	0.435	28.614
5-2 LCL	5	Male		0		1/13/2014	1	5/2/2014	9.7	2	1.007	-83.32
5-2 iPSC	5	Male	1/13/2014	4	11	4/8/2014	1	5/2/2014	9.5	2	0.374	42.367
6-2 LCL	6	Male		0		9/4/2013	1	5/2/2014	9.9	2	1.015	-76.701
6-2 iPSC	6	Male	9/2/2013	3	11	12/10/2013	1	5/2/2014	8.6	2	0.378	32.693
1-3 LCL	1	Male		0		9/4/2013	2	5/2/2014	7.7	3	1.019	-80.293
1-3 iPSC	1	Male	9/2/2013	3	11	12/10/2013	2	5/2/2014	7.7	3	0.38	37.563
2-3 LCL	2	Female		0		9/4/2013	2	5/2/2014	7.8	3	1.01	-83.593
2-3 iPSC	2	Female	9/2/2013	3	11	12/10/2013	2	5/2/2014	8	3	0.398	32.269
3-3 LCL	3	Female		0		1/13/2014	2	5/2/2014	7.8	3	1.019	-81.094
3-3 iPSC	3	Female	1/13/2014	4	11	4/8/2014	2	5/2/2014	7.9	3	0.365	38.609
4-3 LCL	4	Female		0		1/13/2014	2	5/2/2014	7.8	3	0.986	-80.104
4-3 iPSC	4	Female	1/13/2014	4	11	4/8/2014	2	5/2/2014	8	3	0.388	38.995
5-3 LCL	5	Male		0		1/13/2014	2	5/2/2014	8	3	1	-84.215
5-3 iPSC	5	Male	1/13/2014	4	11	4/8/2014	2	5/2/2014	8.1	3	0.373	40.1
6-3 LCL	6	Male		0		1/13/2014	2	5/2/2014	7.9	3	0.981	-81.691
6-3 iPSC	6	Male	1/13/2014	4	11	4/8/2014	2	5/2/2014	7.6	3	0.365	29.021

Table 2.2: Sample information: Includes the following information for all lines: sample ID, gender, reprogramming, extraction, and array batch assignments and dates, RIN scores, passage of RNA collection for iPSCs, and PluriTest scores for all samples.

Because the regulation of a large percentage of genes was affected by reprogramming (67% of tested genes), we asked whether gene expression patterns specific to the donor individual were recovered in the process. We addressed this question using two approaches. First, we evaluated the overall degree of similarity across cell lines from the same donor by considering summaries of the gene expression phenotypes using clustering analysis and PCA. The rationale

for collapsing our gene-specific expression data and considering overall summaries is that complex phenotypes can often be the result of a large combination of genotype contributions and we are interested to learn whether the overall data from cell lines exhibits a clear signature of the donor. In our second approach, we focused on gene specific patterns by partitioning the variance in expression levels for individual genes and testing for differences between the entire distributions of gene expression levels across cell lines. In this approach we are considering expression patterns of individual genes as independent data points. The rationale for the gene-specific approach is that studies of the genetic basis for regulatory variation (such as eQTL mapping studies) nearly always consider the expression phenotypes of individual genes and we are interested to learn the extent to which the effect of donor genotype on gene expression levels can be studied using a given cell model.

To evaluate overall clustering properties in the expression data from the two cell types, we performed hierarchical clustering analysis and PCA. As we performed these analyses, we consistently observed that data from the second iPSC line of individual 4 (line marked as 4-2 in our figures) accounts for a disproportionate amount of variance (Fig. 5). This individual is a clear outlier and its iPSC is associated with the lowest PluriScore in our study (Table 2). We have excluded the data from this individual from subsequent analyses. Importantly, we have confirmed that our conclusions our robust with respect to this decision.

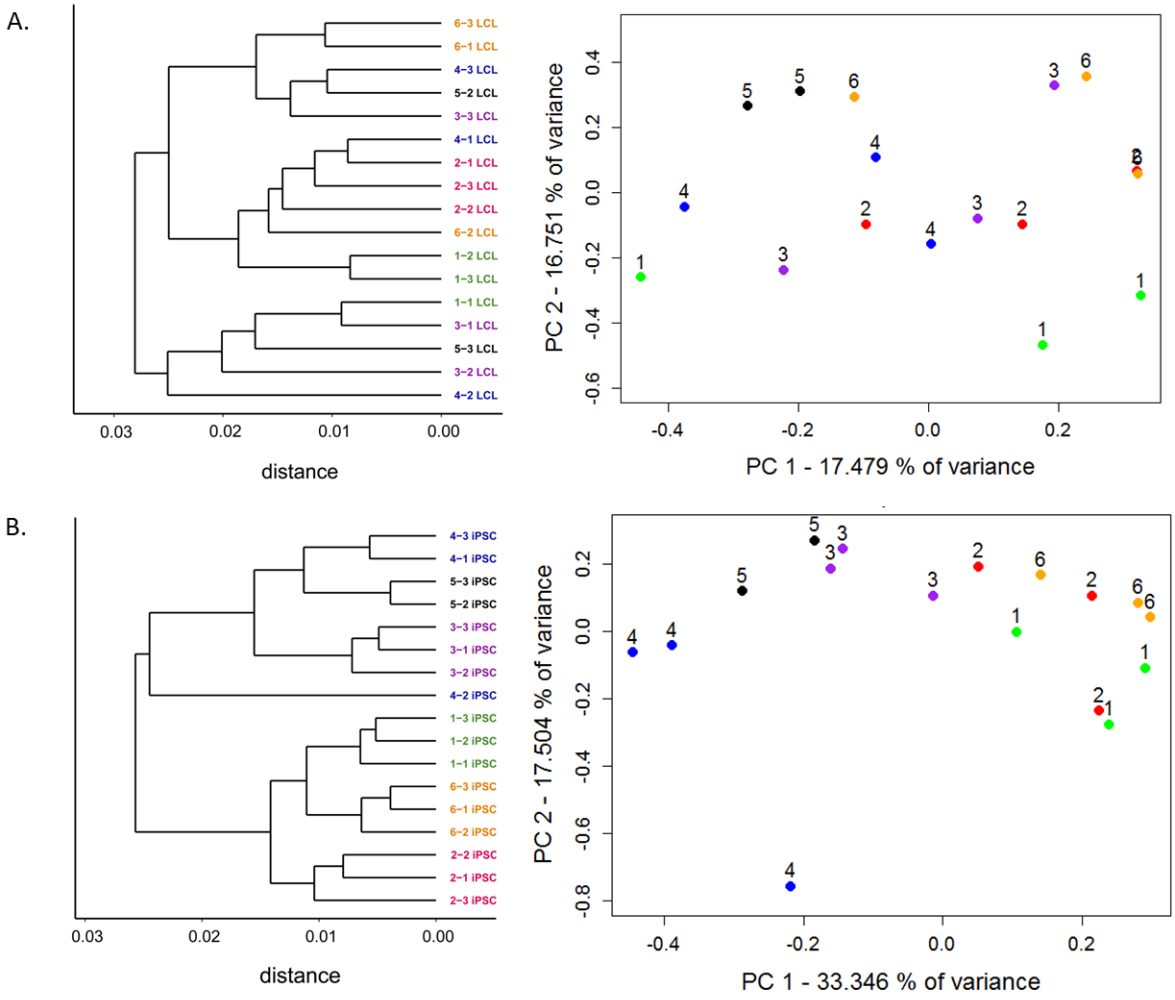


Figure 2.5: Clustering Analysis: A. Results from hierarchical clustering analysis of microarray gene expression and expression data projections on principal components axes 1 and 2 from cycle 7 LCLs and B. iPSCs. Includes data from all lines.

Using data from all 12,243 genes detected as expressed, mature LCLs fail to consistently cluster by the individual from whom they were initially derived, in accordance with our previous observations (Fig. 5A, 6A, 7A). Data from the corresponding iPSC lines, however, cluster by the individual of origin, indicating a large degree of recovery of donor gene expression patterns

(Fig. 5B, 6B, and 7B Figs.). Another method to assess overall clustering properties is through the use of principal components analysis. Taking this approach, we found that clustering of the expression data by individual of origin is substantially more pronounced in the iPSCs than in the LCLs (Fig 5 and 6.). Indeed, the average pairwise Euclidean distances of expression data projections on the first two PCs are significantly smaller within cell lines derived from the same individual than those from different individuals for iPSCs ($P < 10^{-15}$), but not for LCLs ($P = 0.13$; Table 3).

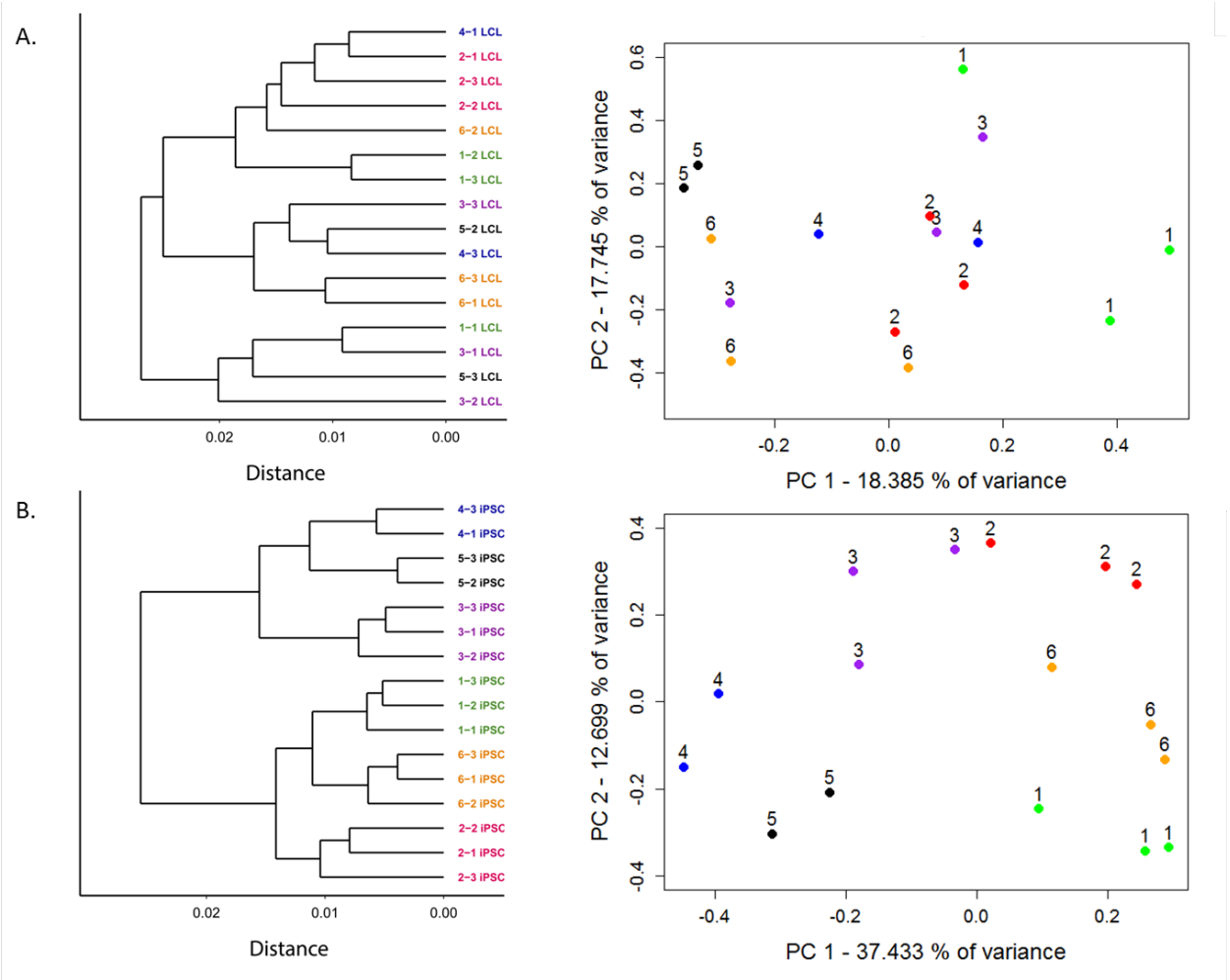


Figure 2.6: Improved clustering properties after reprogramming to iPSCs. A. Results from

hierarchical clustering analysis of microarray gene expression and expression data projections on principal components axes 1 and 2 from cycle 7 LCLs and B. iPSCs.

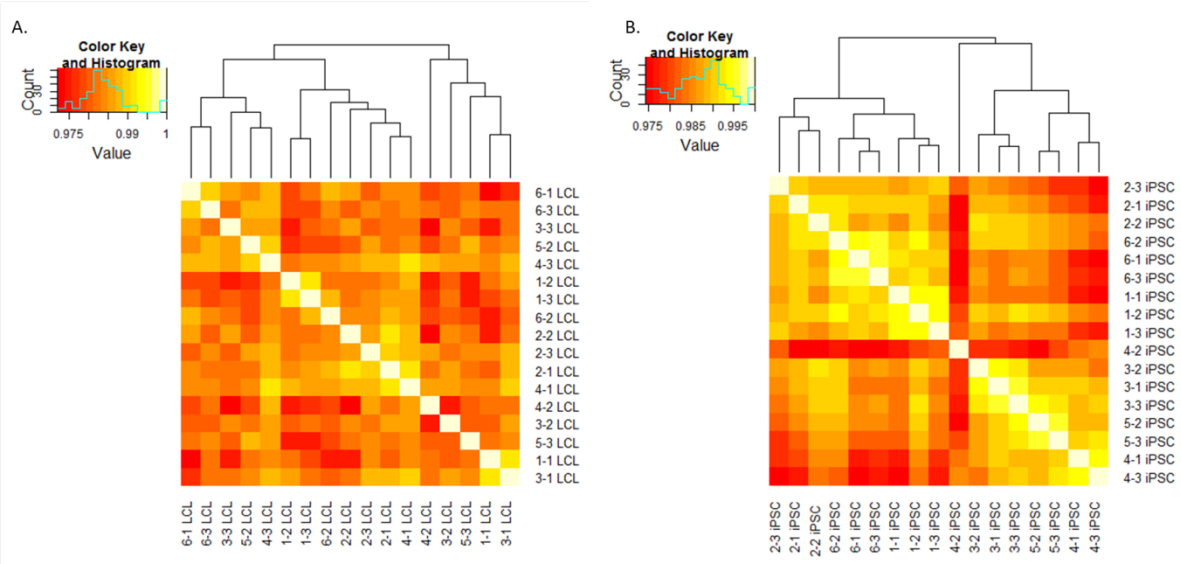


Figure 2.7: Correlation Heatmaps: Heatmap generated from pairwise correlation matrix (pearson product-moment correlation coefficients) for A. LCLs and B. iPSCs. Includes data from all lines.

Cell Type	ED12 within	ED12 between	p-value
Excluding 4-2			
iPSCs	0.1661	0.5158	1.27 e -13
LCLs	0.342	0.4837	1.03 e -01
Including 4-2			
iPSCs	0.2531	0.436	0.024
LCLs	0.3142	0.4588	0.142

Table 2.3: Euclidean distances between clusters in principal components analysis: Euclidean distances of sample projections on the first two principle component axes within and between individuals averaged over all samples for each cell type, demonstrating significantly lower

within-individual distances compared to between-individual distances in iPSCs but not LCLs, regardless of outlier inclusion status.

To estimate the magnitude of the donor effect on gene expression patterns in LCLs and iPSCs, we compared the pairwise correlations of expression data from cell lines derived from the same donor to pairwise correlations of data from cell lines derived from different individuals (Fig 8.). On average, both within- and between-donor correlation coefficients are significantly higher in iPSCs than in the LCLs they were initially derived from ($p < 10^{-4}$ and $p < 10^{-8}$ for within- and between-donor correlations, respectively). In other words, regardless of the individual of origin, we observed less variation in gene expression between iPSCs than LCLs. Yet, though iPSCs harbor less variation overall, the proportion of variation in gene expression that is explained by donor is significantly higher in the iPSCs compared with the LCLs ($P < 10^{-15}$). Indeed, using a single factor ANOVA, we estimate that donor explains, on average, 24.5% of the variance in gene expression in iPSCs but only 6.9% in LCLs.

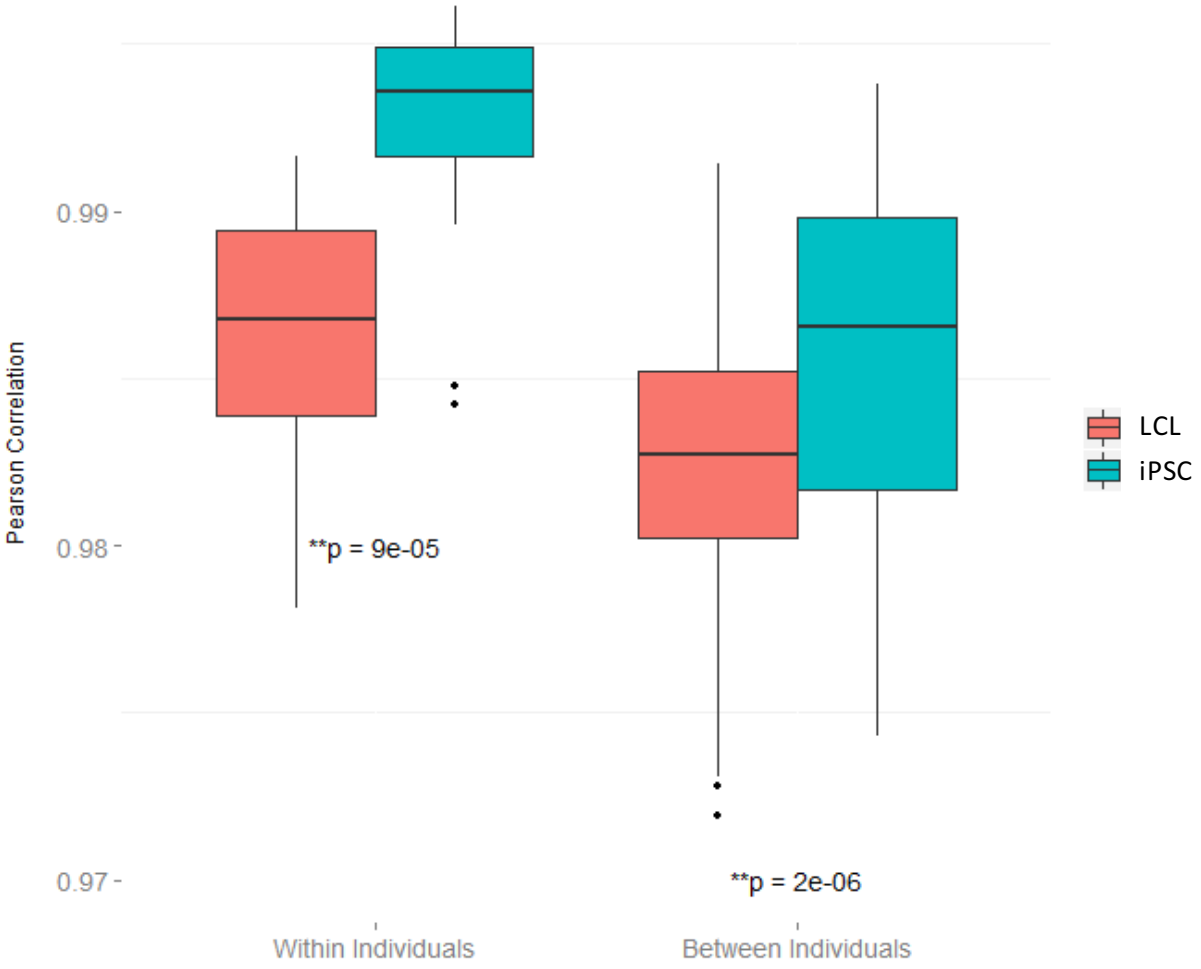


Figure 2.8: Within and between individual expression correlation: Pairwise Pearson correlation coefficients for gene expression data from lines derived from the same individual and across different individuals for both cell types. iPSCs demonstrate increased correlation both within and across individuals compared with LCLs. Includes data from all lines.

In addition to within-donor correlations, we were specifically interested in identifying genes that were highly variable across donors. We thus proceeded by considering the ratio of

between- to within-individual variation in gene expression levels in the two cell types. On average, we found a significantly higher ratio of between-to-within individual variance in gene expression levels in iPSCs compared with data from the LCLs ($P < 10^{-15}$; recall that in this analysis we consider expression patterns of individual genes as independent data points), despite significantly higher overall variance in LCL gene expression ($P < 10^{-14}$; Fig. 9, 10, 11). We identified 1,620 genes whose expression levels were significantly associated with donor in iPSCs (single factor ANOVA FDR < 0.05; see Fig. 12 for histogram of p-values) but only 77 such genes in LCLs.

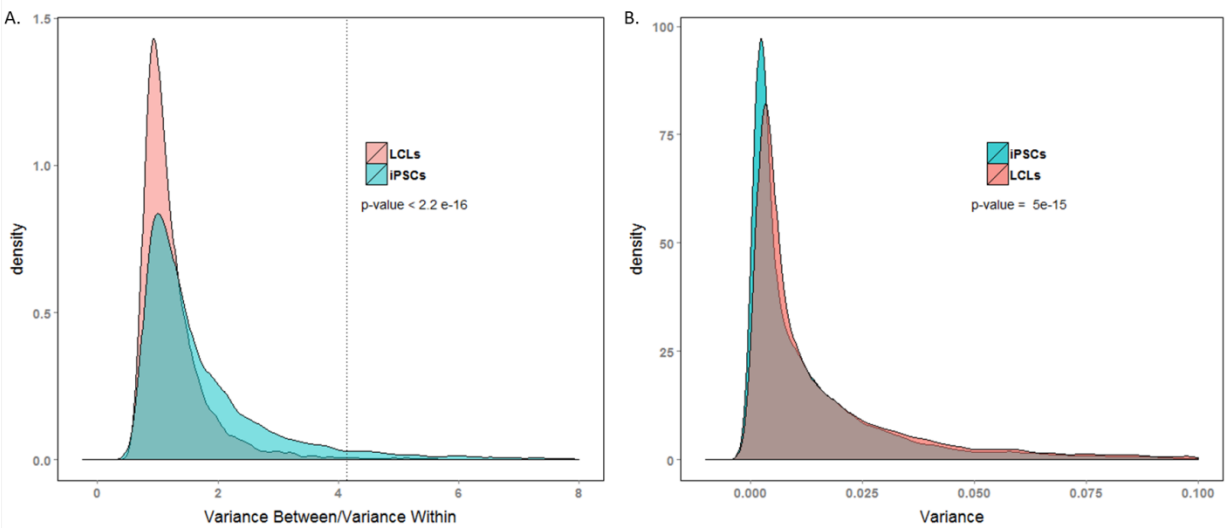


Figure 2.9: Comparison of ability to detect inter-individual gene expression variation. A. Density plot of between donor variance to within donor variance in gene expression for all expressed genes in iPSCs and LCLs. The dotted line indicates the threshold ratio corresponding with significant association between gene expression and donor. X-axis was truncated at 8.0; 0.8% of the data are not plotted here for visualization purposes. B. Density plot of total variance in LCLs and iPSCs. X-axis was truncated at 0.1; 3.7% of the data are not plotted here. See S7 Fig for plots including all the data.

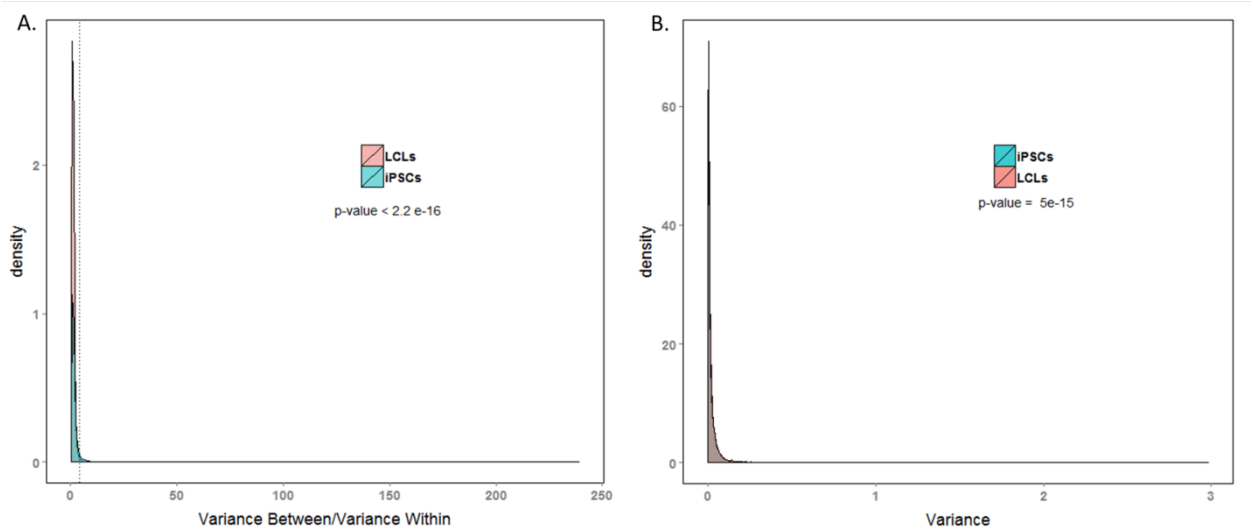


Figure 2.10: Density plots of gene expression variance**:** A. Density plot of between donor variance to within donor variance in gene expression for all expressed genes in iPSCs and LCLs The dotted line indicates the threshold corresponding with significant association between gene expression and individual of origin. All data are plotted; calculations exclude the outlier. B. Density plot of total variance in LCLs and iPSCs. All data are plotted; calculations exclude the outlier.

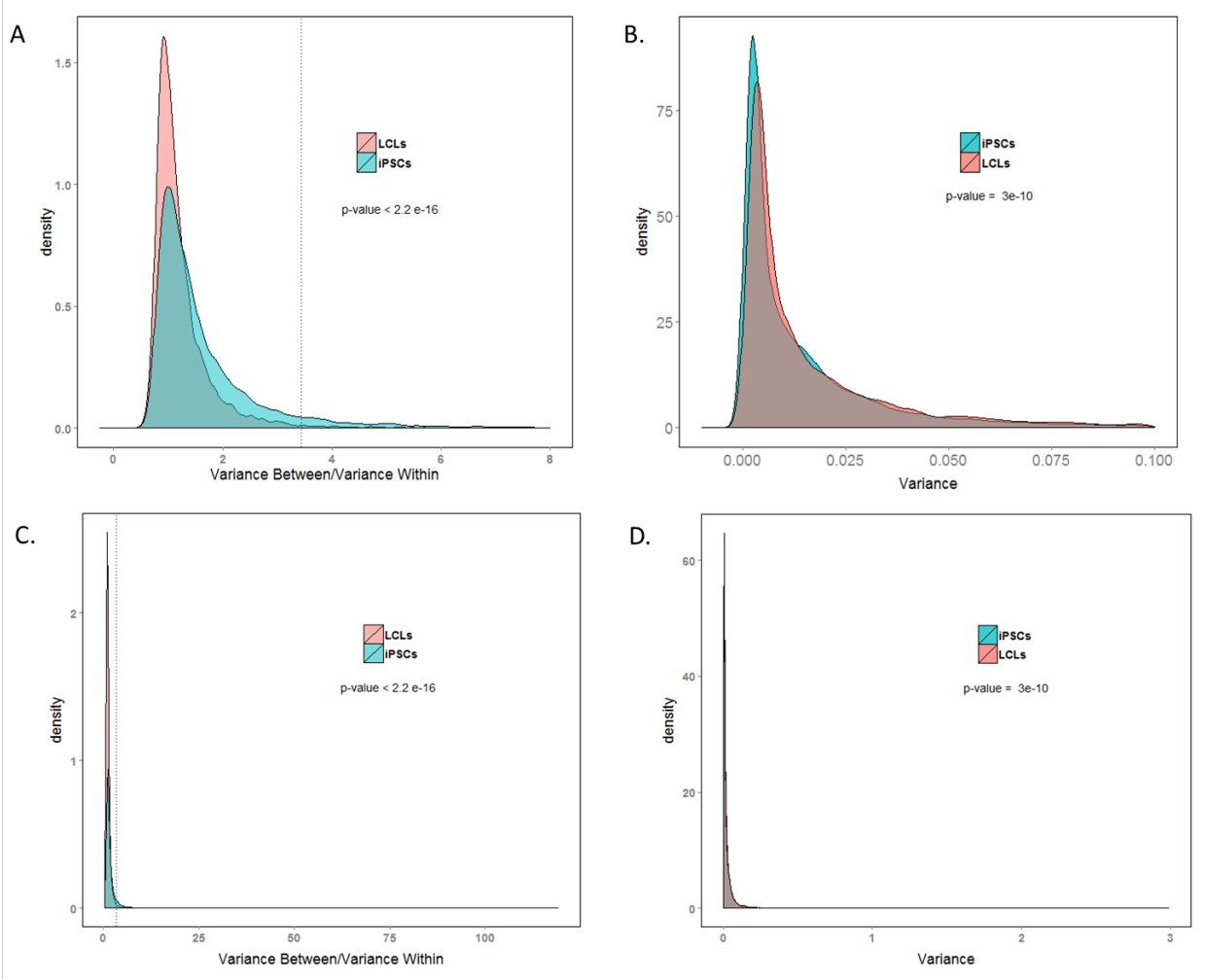


Figure 2.11: Density plots of gene expression variance in all lines: A. Density plot of between donor variance to within donor variance in gene expression for all expressed genes in iPSCs and LCLs. The dotted line indicates the threshold corresponding with significant association between gene expression and individual of origin. X-axis was truncated at 8.0; 0.67% of the data are not plotted here for visualization purposes. Calculations include data from all lines. B. Density plot of total variance in LCLs and iPSCs. X-axis was truncated at 0.1; 4.0% of the data are not plotted here. Calculations include data from all lines. C. Density plot of between donor variance to within donor variance in gene expression for all expressed genes in iPSCs and LCLs. The dotted line indicates the threshold corresponding with significant association between

gene expression and individual of origin. Calculations include data from all lines and all data are plotted. D. Density plot of total variance in LCLs and iPSCs. Calculations include data from all lines and all data are plotted

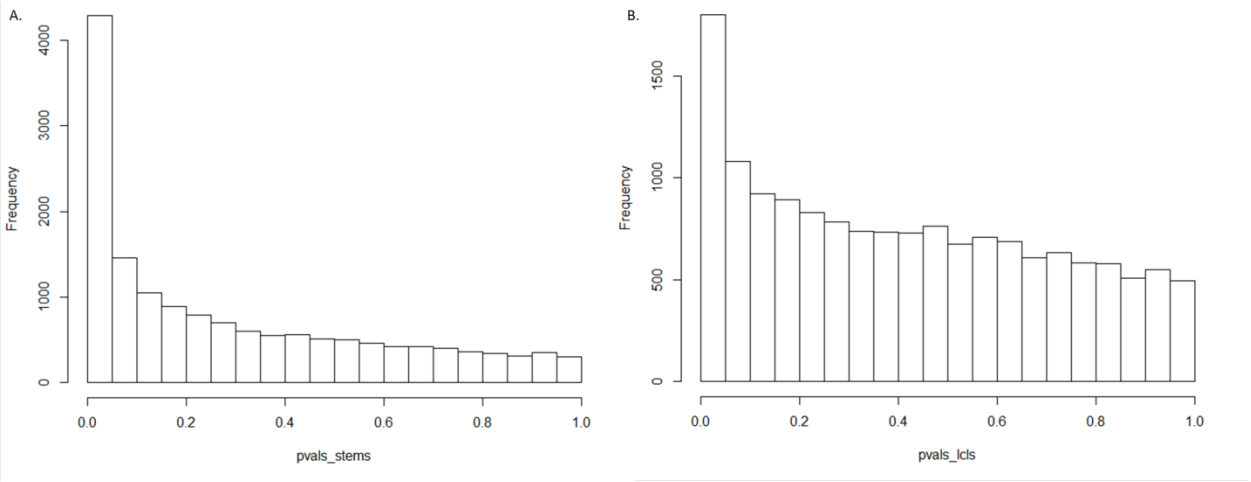


Figure 2.12: Unadjusted p-values for donor effect: Histogram of unadjusted p-values from ANOVA F-test across the factor individual of origin for A. iPSCs and B. LCLs. Includes data from all lines.

Functional Relevance of Highly Variable Genes

We tested for enrichment of functional annotation related to tissue-expression, disease involvement, and biological process (using the online database Lynx [82]) among genes whose expression levels are significantly associated with donor. While these results do not shed much light on the functional importance of these gene sets, we note that genes exhibiting high individual variation in iPSCs are enriched in genes expressed in embryonic tissue while those in

LCLs are not significantly enriched in any functional category we tested. The complete set of results for tissue, disease, and biological process enrichment is available in Tables 4 & 5.

Feature ID	Name	Data Source	In Query	In Test Set	P Value	Bayes
LCL: GO Hierarchy - Biological Process Enrichment:						
GO:0002829	negative regulation of type 2 immune response	GO Hierarchy	2	4	0.001	3.091
GO:0050777	negative regulation of immune response	GO Hierarchy	3	39	0.002	1.753
GO:0019363	pyridine nucleotide biosynthetic process	GO Hierarchy	2	10	0.003	1.647
GO:0019359	nicotinamide nucleotide biosynthetic process	GO Hierarchy	2	10	0.003	1.647
GO:0009435	NAD biosynthetic process	GO Hierarchy	2	10	0.003	1.647
GO:0072525	pyridine-containing compound biosynthetic process	GO Hierarchy	2	12	0.003	1.338
GO:0045191	regulation of isotype switching	GO Hierarchy	2	13	0.004	1.201
GO:0019674	NAD metabolic process	GO Hierarchy	2	15	0.005	0.955
GO:0002828	regulation of type 2 immune response	GO Hierarchy	2	17	0.006	0.738
GO:0042537	benzene-containing compound metabolic process	GO Hierarchy	2	18	0.007	0.639

Table 2.4: Lynx enrichment analysis in LCLs: Enrichment results from selected GO categories for genes with a significant donor effect in LCLs. Data downloaded from the online database Lynx.

Feature ID	Name	Data Source	In Query	In Test Set	P Value	Bayes
iPSC Tissue Enrichment (FDR = 0.05)						
T59	embryonictissue	NCBI Unigene	1363	10323	1.07E-61	139.855
T6	normal	NCBI Unigene	1530	12998	3.13E-55	124.195
T8	adult	NCBI Unigene	1517	12789	3.49E-54	121.461
T12	brain	NCBI Unigene	1473	12165	7.77E-51	113.409
T19	lung	NCBI Unigene	1408	11277	1.17E-49	110.598
T48	fetus	NCBI Unigene	1468	12128	1.43E-49	110.231
T41	germcelltumor	NCBI Unigene	1376	10874	3.6E-49	109.286
T15	eye	NCBI Unigene	1352	10794	3.7E-42	92.93
T27	prostate	NCBI Unigene	1285	10018	2.84E-41	90.897
T17	kidney	NCBI Unigene	1351	10817	4.07E-41	90.444
iPSC Disease Enrichment (FDR = 0.05)						
C3262	tumor	Cancer Gene Index [CGI]	393	2855	0.0000682	13.913
C2955	colorectal carcinomas	Cancer Gene Index [CGI]	110	634	0.007	7.915
C4872	breast cancer	Cancer Gene Index [CGI]	193	1330	0.034	6.454

Table 2.5: Lynx enrichment analysis in iPSCs: Enrichment results from selected GO categories for genes with a significant donor effect in iPSCs. Data downloaded from the online database Lynx.

Finally, we considered the relevance of our findings with respect to previously published eQTL studies in LCLs. Our sample of 6 individuals is too small to allow identification of eQTLs. As an alternative, we compared individual variation in expression levels between genes previously associated with an eQTL in LCLs [15], and genes for which an eQTL was not identified. To do so, we randomly selected data from one biological replicate (one LCL and its corresponding iPSC) from each individual.

In both LCLs and iPSCs, the average coefficients of expression variation were significantly higher in genes previously associated with eQTLs than in genes for which eQTLs were not identified ($P < 10^{-10}$ and $P = 0.009$, for LCLs and iPSCs, respectively; Fig 13, 14, 15). As expected (given that these eQTLs were originally observed in LCLs, and that LCLs have greater overall variation), the coefficients of variation are significantly higher in eQTL-associated genes in LCLs than iPSCs ($P < 10^{-9}$).

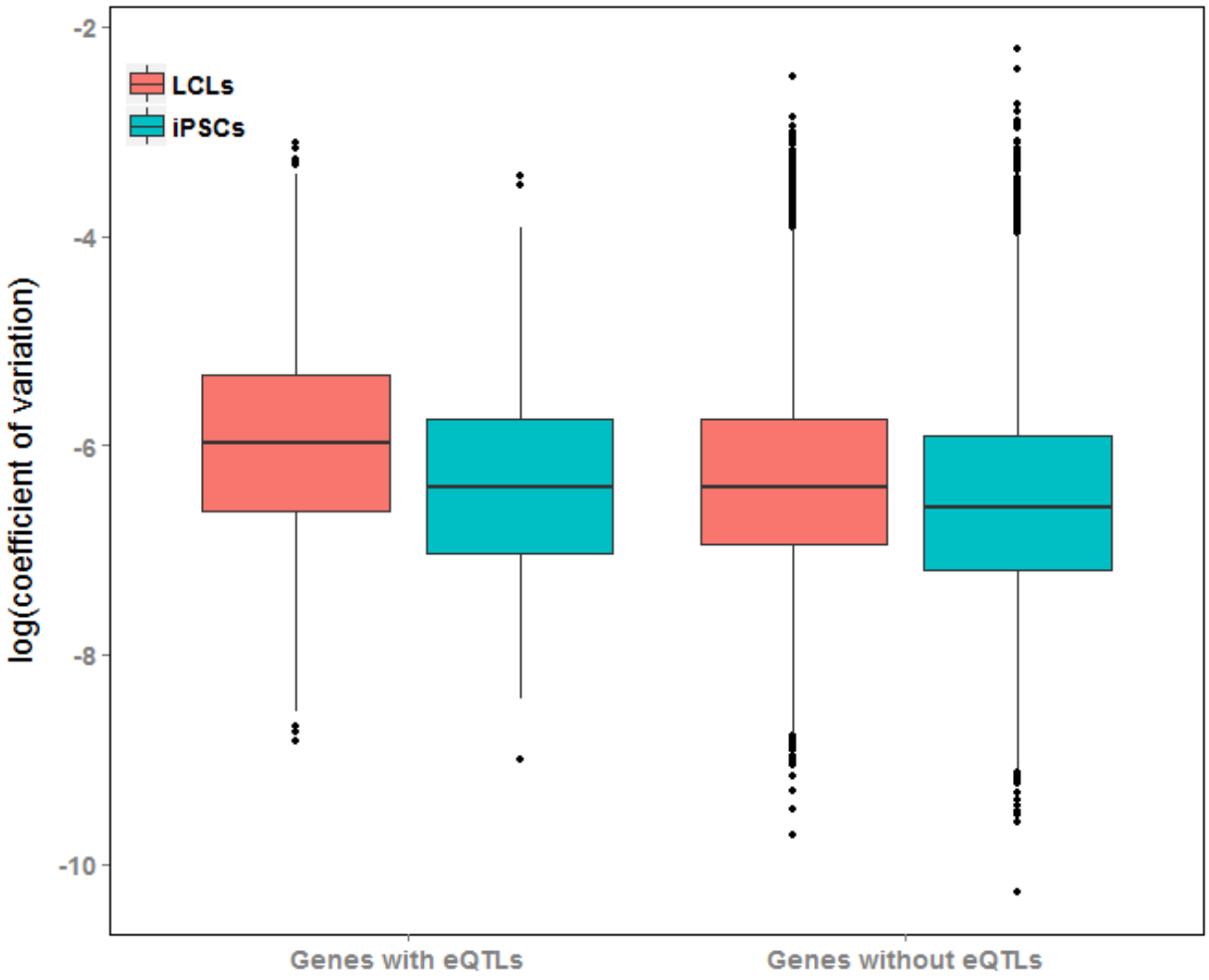


Figure 2.13: Genes with eQTLs are highly variable in both cell types. Boxplot of coefficients of variation of gene expression in genes with and without eQTLs previously identified in LCLs [6], plotted for LCLs and iPSCs. Both cell types exhibit higher c.v,s in genes previously associated with eQTLs than in genes for which eQTLs were not identified $P < 10^{-10}$ and $P = 0.009$, for LCLs and iPSCs, respectively

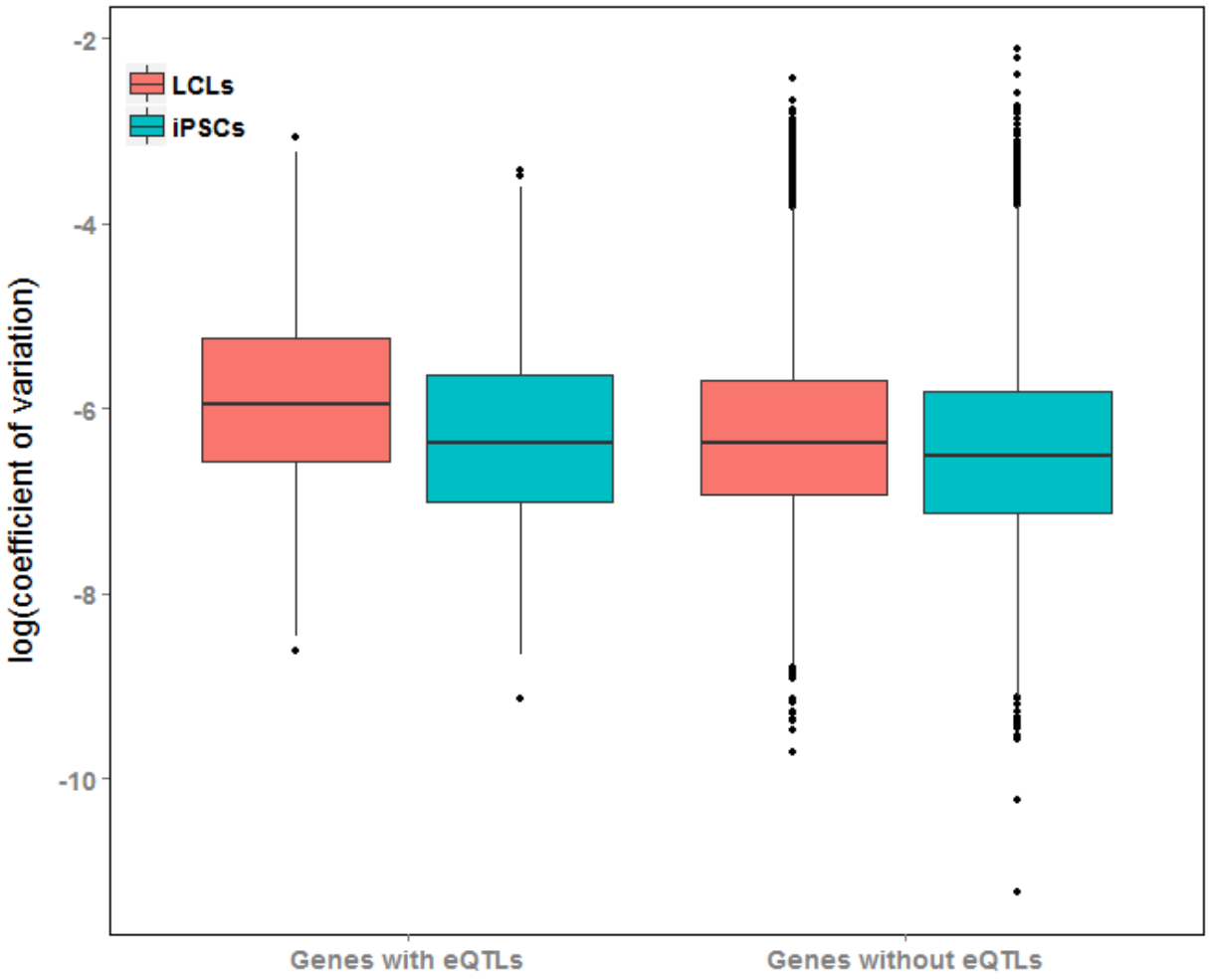


Figure 2.14: Coefficient of variation in genes with and without eQTLs: Boxplot of coefficients of variation of gene expression in genes with and without eQTLs previously identified in LCLs plotted for LCLs ($P < 10^{-10}$) and iPSCs ($P = 0.01$). Includes data from all lines.

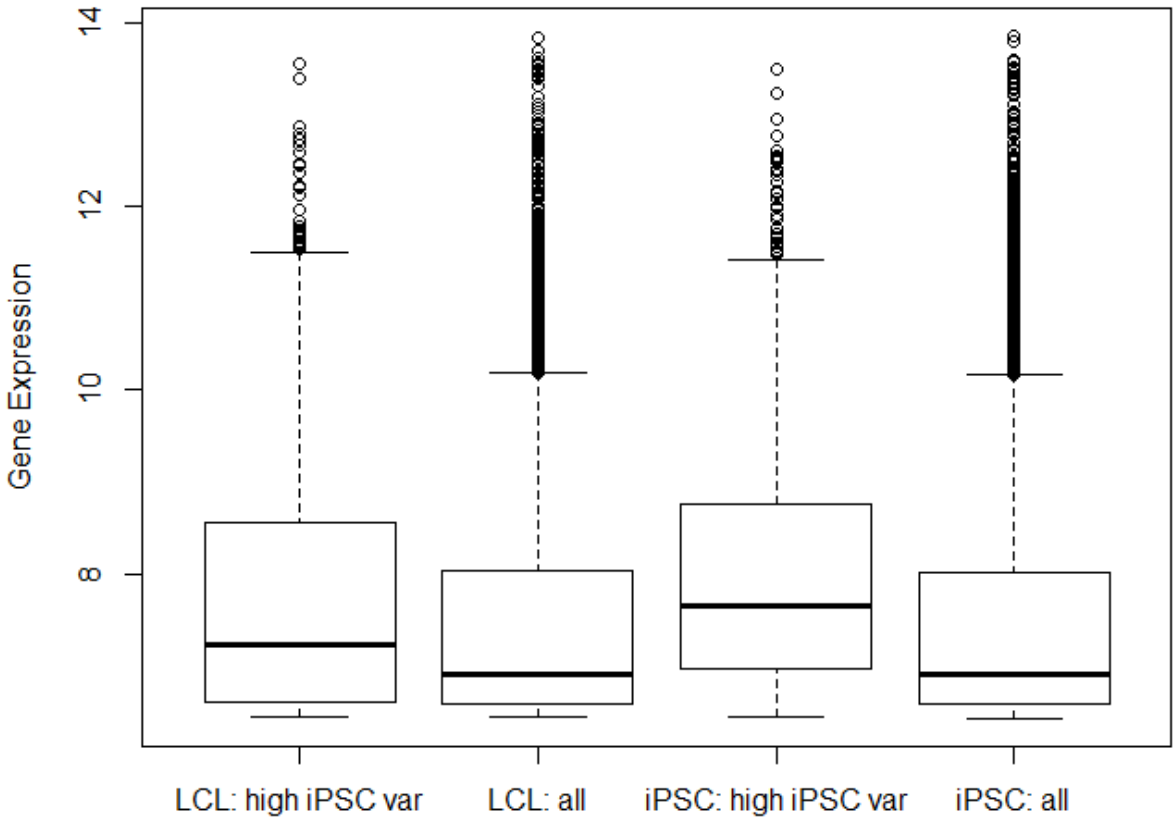


Figure 2.15: Expression in genes with a donor effect in iPSCs: Mean gene expression levels for genes for which a donor effect was detected in iPSCs compared to all genes. Genes with a strong iPSC donor effect are expressed in LCLs, in fact with a higher mean expression value than the genome-wide average ($P < 10^{-15}$). These genes exhibit significantly higher expression than the genome-wide average in iPSCs as well ($P < 10^{-15}$)

2.3 Discussion

The utility of renewable cell lines for population genetics studies and as models of complex disease depends on the preservation of the genotype-phenotype relationship in the cell line. Thus, the most useful cell line models would retain a strong influence of individual of origin on their phenotypes, including molecular properties such as gene regulatory patterns. In previous work, we reported that freeze-thaw cycling of LCLs, a standard and required practice in long-term cell line maintenance, reduces the effect of donor on the cell line's gene expression profile [25]. In fact, we have found that whole-genome gene expression profiles from LCLs that were generated from different individuals are typically as similar to each other as data from independently established replicates of LCLs from the same individual. We suggested that LCLs that have experienced one or more freeze-thaw cycles may be clonally selected for, resulting in a convergent "LCL regulatory phenotype", which has an advantage growing in culture but masks many of the original gene expression differences between the donor individuals.

Apart from the concern regarding the loss of much of the gene regulatory variation between donors, an intrinsic limitation of the LCL model system is that it theoretically represents the biology of only one primary cell type, B cells. As might be expected, all existing collections of renewable cell lines from human population samples include only easily accessible primary tissues such as blood cells, adipocytes, and skin fibroblasts. Many cell types affected by disease, for example cardiomyocytes, hepatocytes, and neurons, cannot be directly studied using existing human cell line panels. In order to study variation in the most relevant phenotypes and

disease processes, we need access to population samples that model additional cellular contexts.

The advent of iPSC technology may have provided the answer. It is now possible to establish renewable iPSC lines from population samples and differentiate them to multiple different cell types for which large collections are currently unavailable. One can establish iPSCs from newly collected fibroblasts or fresh blood samples, but a most attractive possibility is to generate iPSC panels from the already available extensive collections of human LCLs. We thus asked whether individual variation in gene expression levels can be restored by reprograming LCLs into iPSC lines.

Recovery of Individual Variation.

We have shown that not only does the iPSC model exhibit a strong effect of donor on overall gene expression, but in fact the process of reprogramming highly manipulated immortalized cell lines to iPSCs recovers the inter-individual variation in gene expression lost during long term cell line maintenance.

The stronger clustering properties of expression data from iPSCs compared to LCLs suggest that iPSCs are better able to capture donor differences in gene regulation than LCLs. We could detect no significant difference between Euclidean distances within- and across- individuals in the projections of gene expression data from LCLs on the first two principal components of variation, indicating that donor is not a significant global source of gene expression variation in the LCL model. This observation is consistent with our previous findings [25]. In contrast, we

observed a dramatic increase in the number of genes whose expression was significantly associated with donor in iPSCs, and a higher average variance in expression explained by individual of origin in the iPSCs compared with LCLs. These findings indicate that iPSCs reprogrammed from LCLs are a suitable model for studies of donor differences in gene regulation and genotype-phenotype interactions.

Our work does not provide direct evidence for a mechanism by which iPSCs regain the effect of donor on gene expression. Previously [25] we have hypothesized that the loss of individual variation in gene expression levels in LCLs is due to selection in culture for the fastest growing LCLs. We suggested that this selection results in a convergence to a gene regulatory profile that is common for all mature LCL cultures. In a recent study [56] we have found that DNA methylation profiles in iPSCs reprogrammed from different somatic cell types of the same individuals are practically identical, while we observed hundred of thousands of methylation differences between the precursor somatic cells. Global reprograming of the epigenetic landscape in iPSC lines could potentially be the reason that the LCL gene regulatory signature has been largely replaced by a new regulatory program, which no longer reflects the selection pressures relevant to LCL culturing.

We note that despite their diminished ability to reflect donor differences, attempts to identify instances of genetic regulation of LCL gene expression in mature cell lines have been considered largely successful. Indeed, here we report that genes previously identified as associated with an eQTL in LCLs exhibit higher variance in mature LCLs than those without one. However, our observations suggest that, for future eQTL mapping studies, iPSCs may be a better system than LCLs. While 28.6% of genes with a significant donor effect in LCLs are associated with a

previously identified eQTL in LCLs, only 5.9% of genes with a significant donor effect in iPSCs are associated with such an eQTL. Although expressed in LCLs, often at appreciable levels (S11 Fig.), the majority (>95%) of genes with a strong donor effect in iPSCs do not show such an effect in LCLs. Put together, these observations support our assertion that iPSC can be a better model than LCLs for detecting eQTLs, and more generally, for studies of inter-individual differences in gene regulation.

Technical Noise Associated With Reprogramming.

In any cell model, it is important to consider the magnitude of noise introduced by cell culture relative to biological signal. We note a substantial decrease in within-individual expression correlations for a cell line with a low PluriScore, indicating that we should perhaps reconsider acceptable scores for studies of individual phenotypic variation. However, other technical considerations do not seem to have a marked effect on overall clustering properties. For example, data from the single iPSC line that retained EBV (individual 3, replicate 2; S3 Fig.) clustered with the other iPSC lines derived from that individual. Additionally, we reprogrammed iPSCs in four groups and collected expression data at varying passages (between passage 11 and 13, S2 Table) without apparent batch effects. Because it is currently unclear which factors significantly affect our ability to detect donor differences, potential sources of noise need to be more systematically studied and appropriately controlled for.

Much of the excitement surrounding iPSCs is based on their ability to differentiate into terminal cell types, providing a renewable substitute for previously inaccessible tissues. Our study does

not provide direct evidence that iPSC-derived differentiated cells will also reflect donor differences, however because the pluripotent state is relatively well-conserved compared to terminal cell types [83; 84], we expect that tissues derived from iPSCs will demonstrate an even stronger donor effect on gene expression. That said, we suggest that this expectation needs to be independently confirmed in each differentiated cell type before they are carried into further studies.

2.4 Methods

Ethics Statement

In this study, blood samples from Research Blood Components were analyzed anonymously. Research Blood Components obtained IRB approval and written informed consent from each donor, giving permission to collect their blood and use or sell it at Research Blood Components's discretion, for research purposes.

Sample Acquisition

Whole blood was collected from six healthy Caucasian donors by Research Blood Components LLC (Brighton, MA) with IRB consent between 2009 and 2010. B-Cell isolation and LCL generation were performed at the University of Chicago as described previously [14]. Between February 2011 and October 2012, each line was thawed, cultured, and re-frozen every three months, for a total of six freeze-thaw cycles prior to use in our study to study the effects of freeze-thaw cycling on gene expression profiles [16]. LCLs were cultured in RPMI with 20% FBS and frozen in Recovery Cell Culture Freezing Media (Life Technologies).

iPSC Generation and Validation

All cell culture was performed at 37°C, 5% CO_2, and atmospheric O_2. From each individual, three biological replicates of LCLs were reprogrammed to iPSCs using a similar method to that described previously [22,23]. LCLs were transfected in four batches between August 2013 and January 2014 (S2 Table). One million cells were transfected with 2 μg of each episomal plasmid encoding OCT3/4, shP53, Lin28, SOX2, L-MYC, KLF4, and GFP (Addgene plasmids 27077, 27078, 27080, 27082 [21]) using the Amaxa transfection program X-005. For more details see:

http://giladlab.uchicago.edu/data/LCL_Reprogramming.pdf. Transfected cells were grown in suspension for a week in hESC media (DMEM/F12 supplemented with 20% KOSR, 0.1mM NEAA, 2mM GlutaMAX, 1% Pen/Strep, 0.1 mM BME, and 12.5 ng/mL human bFGF) supplemented with 0.5mM sodium butyrate between days 2–12 post-nucleofection. After seven days, cells were plated on gelatin-coated plates with CF-1 irradiated mouse embryonic fibroblasts and manually passaged as colonies for at least 10 passages. After day 12, cells were grown in hESC media without sodium butyrate. Media was changed every 48 hours. Cell pellets were collected and stored at -80° C until extraction. One biological replicate from individual five failed to reach passage ten and was excluded from all analyses.

Embryoid body assays were performed following the protocol used by Romero et al [31]. Briefly, embryoid bodies were generated by manual colony detachment and were grown in suspension for seven days on low adherent plates in bFGF-free hESC media. They were then plated on 12 well gelatin-coated plates and grown for another seven days in DMEM-based media. Cells were fixed and stained using antibodies against nestin (1:250 SC-71665, Santa Cruz Biotech), α-smooth muscle actin (1:1500, CBL171, Millipore), alpha-Fetoprotein (1:100, SC-130302, Santa Cruz Biotech), and HNF3β (1:100 SC-6554, Santa Cruz Biotech) to detect ectoderm, mesoderm, and endoderm lineages respectively.

DNA was extracted using ZR-Duet DNA/RNA MiniPrep (Zymo) kits according to the manufacturer's instructions. To assess for the presence of plasmid or EBV genome in iPSCs, PCR was performed using the genomic DNA collected from the iPSCs as template (collected at the same time as expression measurements) with primers designed to amplify the 3' end of the EBNA-1 gene (present in both the EBV genome and all reprogramming plasmids) and NEBNext

High-Fidelity 2X PCR Master Mix. For the sample with detectable EBNA-1, we also performed genomic PCR using primers to amplify a region common to all PXCLE reprogramming plasmids, and primers that amplify the BBRF1/LMP2 gene found only in the EBV genome to determine the source of foreign DNA. Primer sequences are available in S6 Table. Fibroblast DNA containing reprogramming plasmids at 0.02 pg/μL was used as a positive control for the PXCLE and EBNA-1 primer sets. LCL DNA (from YRI lines 18508 and 19238) were used as positive controls for the EBV and EBNA-1 primer sets. Fibroblast DNA was used as a negative control for all primer sets.

RNA was extracted using ZR-Duet DNA/RNA MiniPrep kits according to the manufacturer's instructions with the addition of a DNAse treatment step prior to RNA extraction. cDNA was then synthesized using Maxima First Strand cDNA Synthesis Kit (Thermo-Scientific.) RT-PCR for endogenous transcripts of three pluripotency-related transcription factors was performed for all iPSC lines using SYBR Select master mix (Life Technologies.) Primers sequences are available in S6 Table. Data were analyzed using Viia7 software (Life Technologies). All expression levels were normalized to GAPDH. Expression was measured relative to a randomly selected iPSC line.

Gene Expression Quantification

Cell pellets were obtained from LCLs immediately before transfection and from stable iPSCs after at least ten passages. RNA concentration and quality was estimated using the Agilent 2100 Bioanalyzer. Donor expression profiles were quantified using Illumina HumanHT-12 v4 Expression BeadChip Microarrays by the Functional Genomics Core at University of Chicago. Samples were hybridized across three array batches. Biological replicates from an individual

were assigned to different batches to exclude a relationship between batch and individual. The array data were also used for the PluriTest assay as described previously [24].

Data Processing and Analysis

Raw probe data were filtered for probes whose target transcripts were detected as expressed ($P < 0.05$) in at least two samples. Probes targeting expressed transcripts were then mapped to the hg19 reference genome. We excluded those with a quality score below 37, those that did not map uniquely to an Ensembl gene ID, that spanned an exon-exon junction, or that contained a SNP with MAF < 0.05 in European populations (calculated using 1000 Genomes phase I integrated call sets). After filtering, probe intensities from all samples were background corrected, quantile-normalized, and log-2-transformed using the R package 'lumi' [32]. For genes represented by multiple probes, only the 3' most probe was included in subsequent analyses to represent the most complete transcript. Finally, array batch was corrected for using an empirical Bayes method implemented in the R package 'sva'[25,33]

Differential expression was estimated using a linear model based empirical bayes method implemented in the R package 'limma [26]'. Dendrograms were generated for matrices of pairwise Pearson product-moment correlation coefficients. For principal component analysis, expression data was mean-centered by gene across all individuals. The outlier individual 4–2 was omitted prior to hierarchical clustering analysis and PCA. All analyses, figures, and tables presented in the supplement include data from all individuals. Proportion of variance due to donor was estimated as the adjusted R2 value from a linear model including a term for each individual. Genes with FDR-adjusted p-values < 0.05 from a one-way ANOVA across individuals were classified as significantly associated with donor. eQTL data were downloaded from the

Pritchard group eQTL browser: http://eqtl.uchicago.edu. Functional group enrichment was assessed using the web-based gene annotation database Lynx: http://lynx.ci.uchicago.edu using all expressed genes subjected to our filtering criteria as background.

Accession Numbers

Gene expression data are available at the GEO database, accession #GSE64263.

CHAPTER 3

An iPSC-based Model of Primate Endoderm Development Reveals a Dynamic Conservation Profile

3.1 Introduction

The relative conservation of particular developmental stages across the animal kingdom is a very old controversy that has been reawakened with the advent of new genomics methods. The roots of the debate began to form in 1828 when embryologist Karl von Baer postulated that early stages of development are highly similar across animals and that these constraints become more relaxed as development progresses[85]. Several decades later, the theory of evolution added a new meaning to this observation, eventually leading to the hypothesis that certain critical events in development are restricted to occur in such a highly precise way so as to be inaccessible to evolutionary forces. This general idea of the canalization of essential, complex developmental processes is widely accepted, but which stages of development are affected by this phenomenon is to this day unclear, despite decades of study.

A very popular proposal for the chronological profile of developmental conservation is the hourglass model. This model was developed to reflect the finding that although vertebrates pass through a state of morphological similarity, termed the phylotypic stage, the structural modes of early processes such as gastrulation differs considerably across species[86]. Anatomical studies generally supported early and late developmental divergence, with a conserved

intermediate transition (reviewed by Richardson and Keuck[87]). However, molecular studies of conservation are not as unanimous in their support for the hourglass model. Initial compelling evidence of global expression correlation across Drosophila species did reveal an elevation in cross-species correlation at the mid-stages of development with lower conservation at the initial and late stages[88]. However, others could find no such divergence at early time-points, and instead reported molecular evidence supporting von Baer's model of early conservation[89]. In fact, a recent study of time-course gene expression data in developing embryos across ten phyla revealed a bulge rather than a neck at mid-development and high conservation at the initial stages of development[90]. Yet another study of global transcript age throughout zebrafish development, and its later reanalysis[91], suggested a more complex pattern within the hourglass including a peak in young transcripts at gastrulation[92].

The conservation status of the earliest stages of development are clearly the most contentious, yet we also view them as the most essential for understanding the foundations of divergence and speciation. In particular, we are interested in the role that canalization of early developmental stages may have played in human evolution. Because developmental processes are largely conserved across animals, and ethical considerations render data from early time-points in primate development exceptionally scarce, studies of molecular development are almost exclusively conducted in model organisms. As a result, we know very little about how these processes occur uniquely in humans and essentially nothing about the temporal profile of divergence of early human development from that of other primates.

Endoderm, the inner-most germ layer in the developing vertebrate embryo, gives rise to all of the organs of the digestive and respiratory systems. Although they are internal, variation in structures arising from the endoderm lineage account for many fundamental differences between humans and other primates. Decades of work in comparative anatomy have catalogued coarse differences in anatomy of structures involved in nutrition and speech. These findings are supported by genetic studies, detecting evidence of positive selection in promoter regions of genes related to nutrition[46] and involved in carbohydrate metabolism, fatty acid processing, and vitamin transport[93]. Species differences in risk for diseases associated with endodermal organs have also been appreciated, such as carcinomas and hepatitis progression[94]. It is well appreciated that the most profound differences between humans and other primates are initiated early in the developing animal, but precisely when these differences arise and through what molecular mechanisms remains completely obscure.

Here, we present a comparative time course analysis of global gene expression changes in humans and chimpanzees during early endoderm development. We explore the chronology of expression divergence in early stages of primate development to estimate the effects of canalization and identify categories and features of genes with particular divergence patterns in the endoderm lineage.

3.3 Results

To study expression dynamics during early primate endoderm development, we analyzed genome-wide RNA-seq data from a panel of humans and chimpanzees at four time-points along endoderm development (Figure 3.1A). Including only genes with a known ortholog in

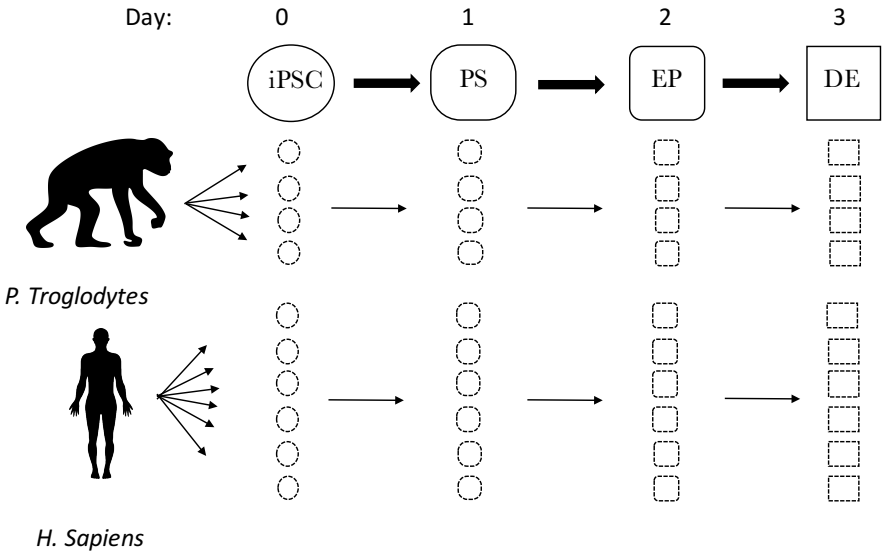
chimpanzees, we quantified expression data for 14,737 genes (see methods for complete inclusion criteria) at all four developmental time-points.

Characterizing Developmental Time-points

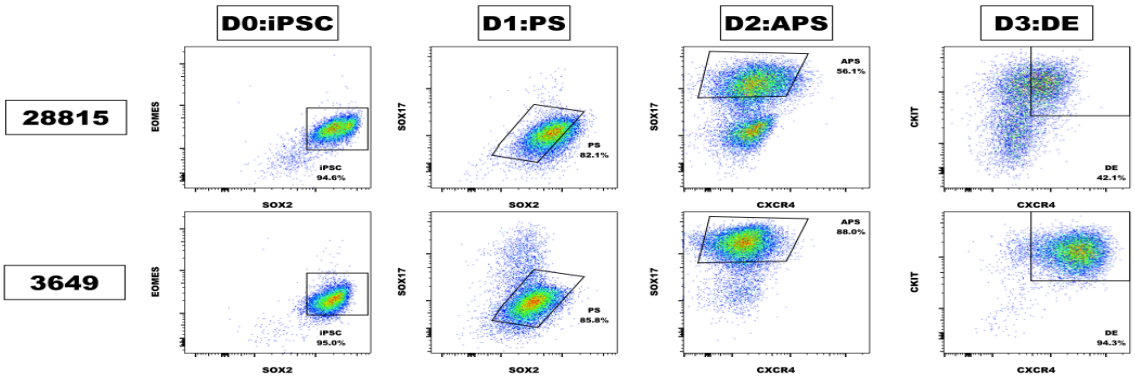
To access the cardinal stages in endoderm development, we targeted critical lineage bifurcations following a directed step-wise differentiation protocol. Through this method, we obtain highly pure iPSCs, primitive streak, endoderm progenitors, and definitive endoderm with comparable purity estimates between the species as assessed by flow cytometry (Figure 3.1B, 3.2). Expression of critical transcription factors for each day was also quantified (Figure 3.1C, Table 3.1) [95-97]. We find that these genes are expressed at high levels in humans and chimpanzees at the relevant day (Figure 3.1C, D). Complete purity information is available in Table 3.2.

Figure 3.1. Study Design and Stage-Specific Markers

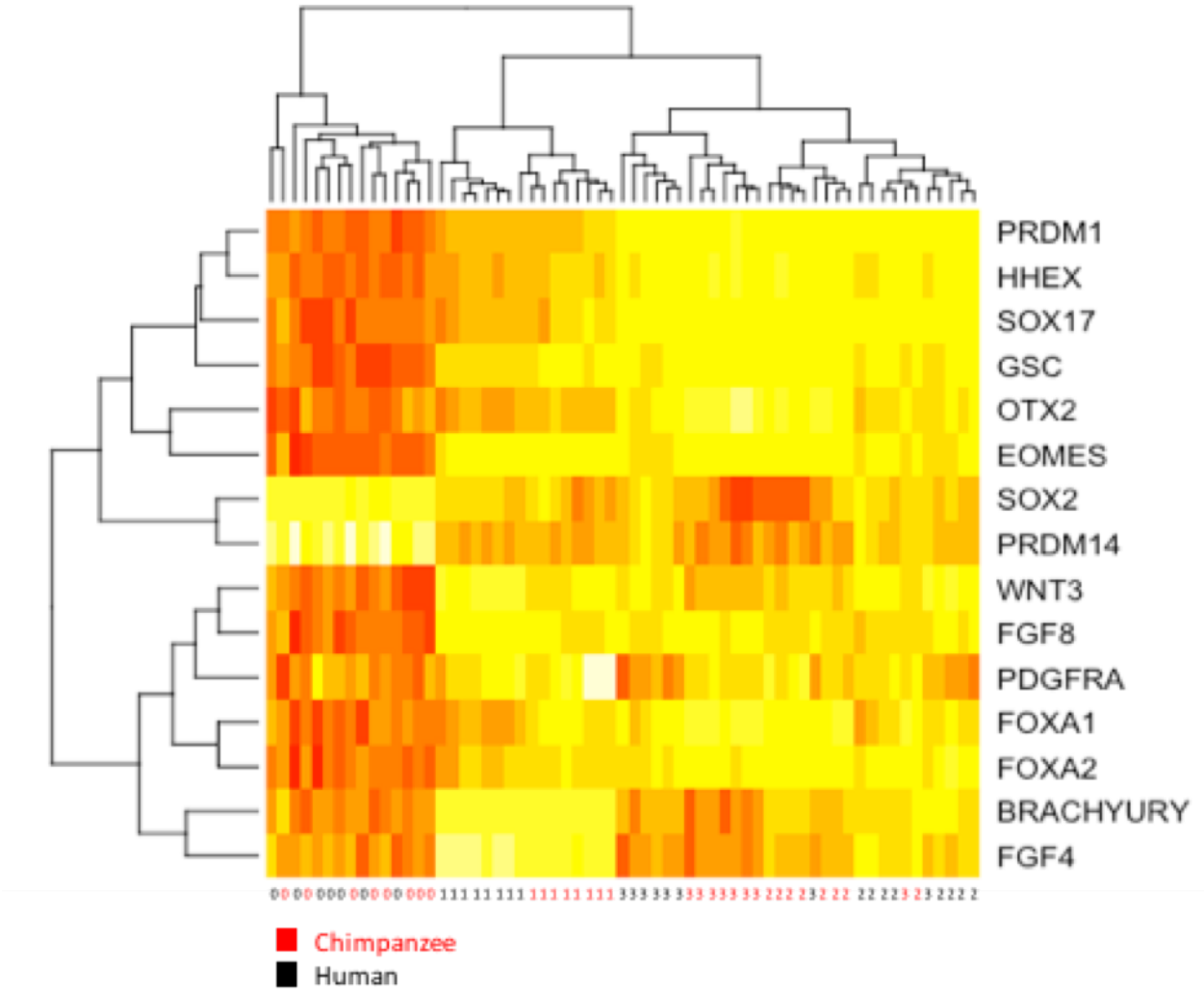
A.

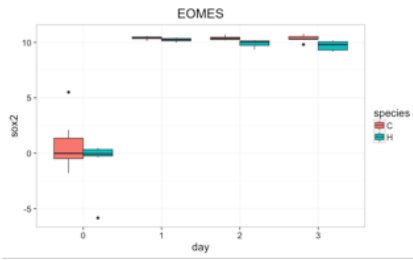


B.

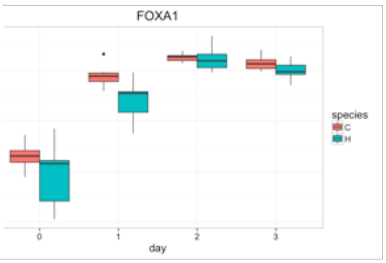


(Figure 3.1 continued) C.

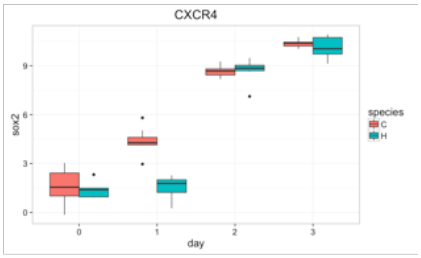




EOMES: Primitive Streak



FOXA1: Anterior Primitive Streak



CXCR4: Definitive Endoderm

Figure 3.1 continued. A. Study design. Four chimpanzees and six humans were studied at four time-points during endoderm development. iPSC: induced pluripotent stem cell, PS: primitive streak, EP: endoderm progenitor, DE: definitive endoderm. B. Representative purity for each time-point as assessed by flow cytometry. 28815: human female, 3649: chimpanzee female. Purity gates are as follows: Day 0: EOMES negative, SOX2 positive, Day 1: SOX2 positive, SOX17 negative, Day 2: SOX17 positive, CXCR4 negative, Day 3: CKIT positive, CXCR4 positive. C. Heatmap of expression of transcription factors known to be highly expressed in one or more stages in our time-course. See Table 3.1. D. Expression level of several critical regulators of endoderm development across the species

Figure 3.2: Example of Gating for Purity Estimates

A.

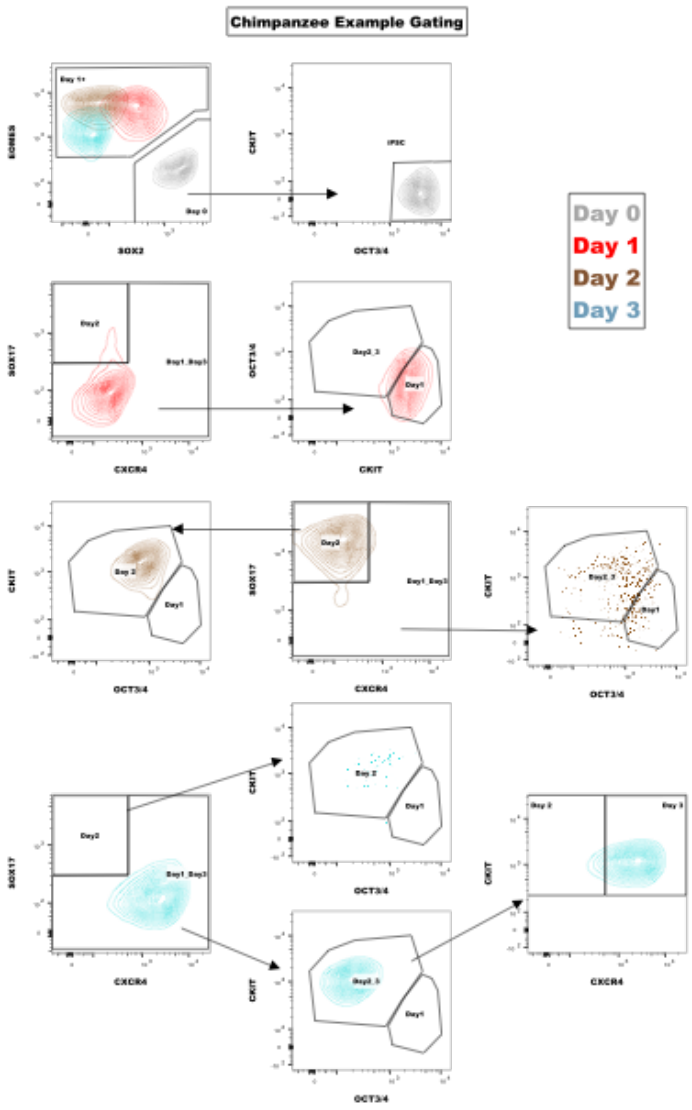


Figure 3.2 continued: B.

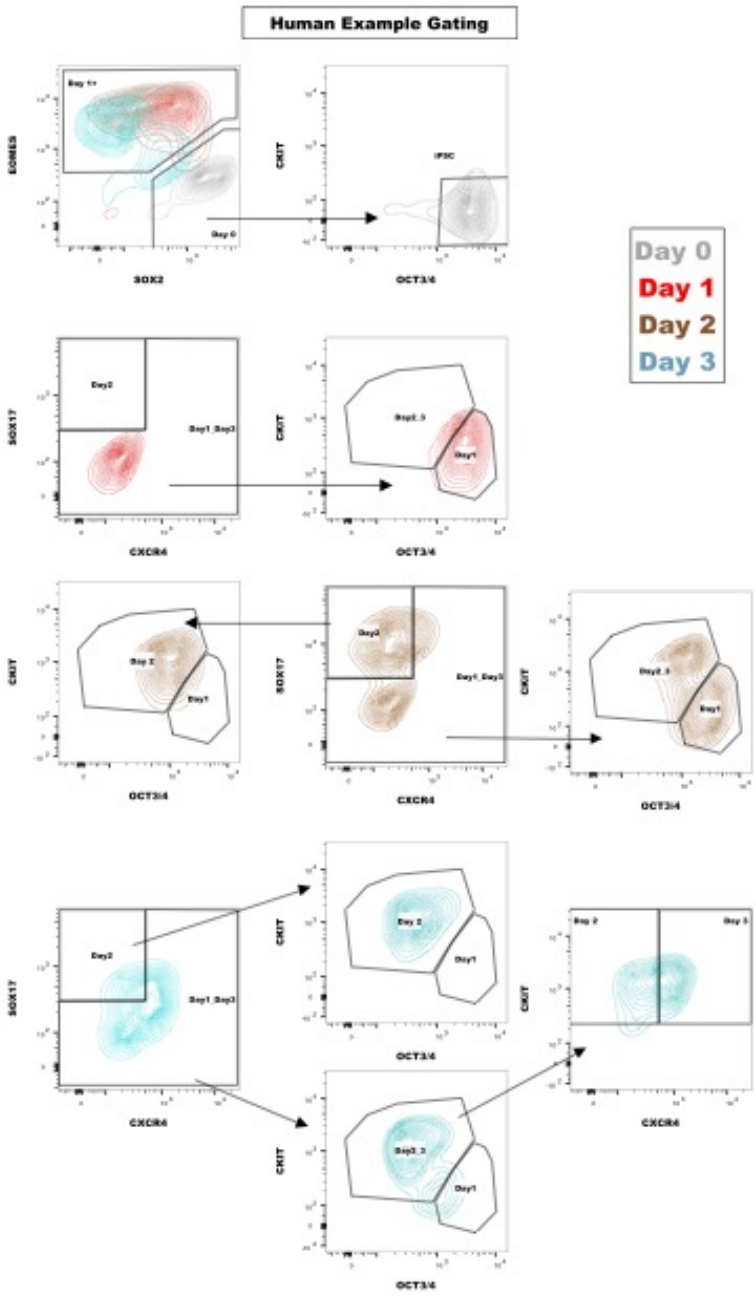


Figure 3.2 continued. A: Gating example for human time-course B. Gating example for chimpanzee time-course. See Table 3.2A for gating scheme..

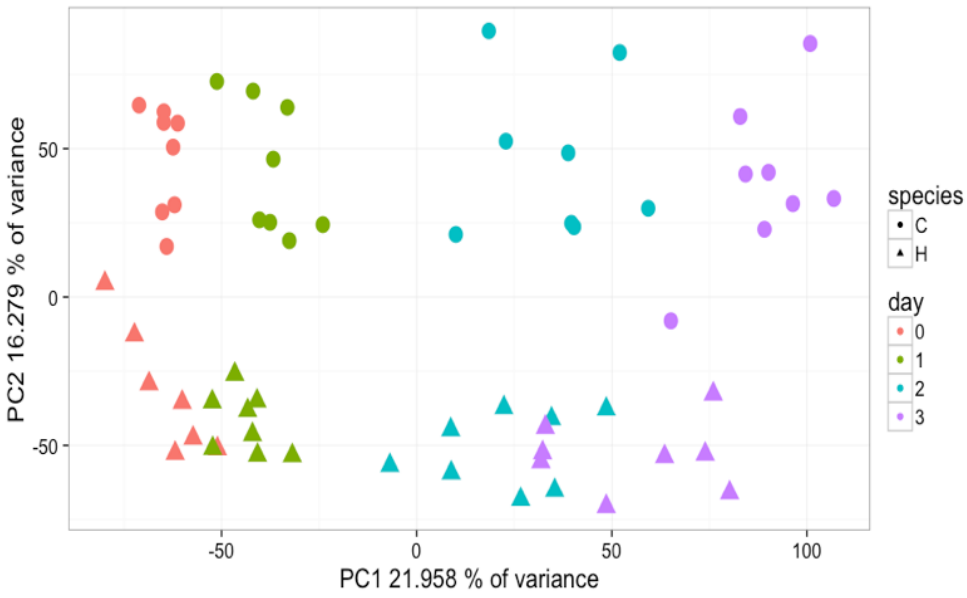
A global survey of our data through principal component analysis indicated that position in the time-course represented the primary source of gene expression variation during the study (explaining 22% of variance), while species was the second strongest driver of variation (explaining 16% of variance). Hierarchical clustering analysis supports this result with time-point primarily driving clustering, followed by species. Both analyses suggest that the final two time-points are closest in expression state. See Figure 3.3.

Pluripotency	Primitive Streak	Endoderm Progenitor	Definitive Endoderm
SOX2	EOMES	GSC	FOXA2
PRDM14	BRACHYURY	FOXA1	OTX2
	PDGFR-A	HHEX	PRDM1
	WNT3	SOX17	SOX17
	FGF4	PRDM1	
	FGF8		

Table 3.1. Transcription factors specific to each developmental stage included in the study (Figure 1C).[95-97]

Figure 3.3: Principal Components Analysis and Hierarchical Clustering

A.



B.

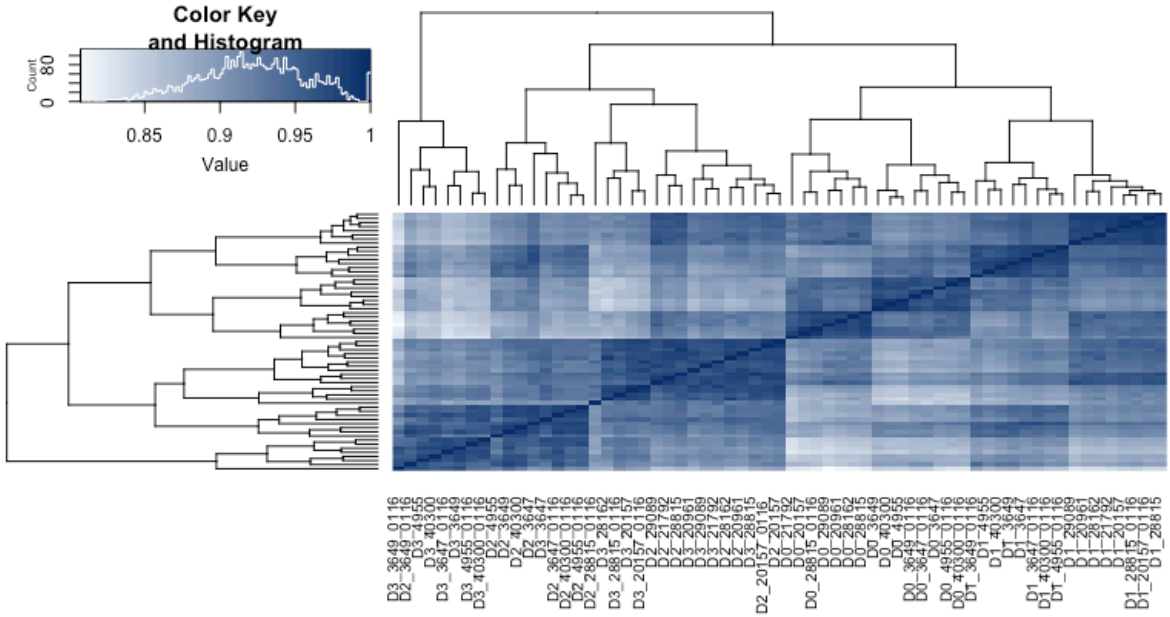


Figure 3.3 continued: A. Global expression CPM measurements projected onto the axes of the first two principal components. Color indicates position in time-course (day). Shape represents species. B. Hierarchical clustering results of pairwise Spearman correlations. Lines ending in _0116 are differentiation replicates.

A.

	Marker					
Day	SOX2	OCT3/4	EOMES	SOX17	CKIT	CXCR4
0	+ High	+ High	-	-	-	-
1	+ Low	+ High	+	-	+ Low	-
2	-	+ Low	+	+	+ High	-
3	-	-	+Low	-	+	+

B.

Sample:	High Confidence Purity	Maximum Purity
Day0_iPSC_001_28126	82.3	-
Day0_iPSC_002_4955	87.2	-
Day0_iPSC_003_29089	86.2	-
Day0_iPSC_004_40300	88.1	-
Day0_iPSC_005_28815	85.7	-
Day0_iPSC_006_3649	95.6	-
Day0_iPSC_007_20157	88.7	-
Day0_iPSC_008_3647	79.9	-
Day1_PS_001_28126	72.96%	83.7%
Day1_PS_002_4955	76.21%	90.0%
Day1_PS_003_29089	78.93%	84.8%
Day1_PS_004_40300	66.41%	84.8%
Day1_PS_005_28815	76.10%	85.4%
Day1_PS_006_3649	73.28%	84.0%
Day1_PS_007_20157	73.02%	92.9%
Day1_PS_008_3647	66.19%	74.3%
Day2_APS_001_28126	62.7%	76.3%
Day2_APS_002_4955	84.9%	92.7%
Day2_APS_003_29089	35.9%	37.0%
Day2_APS_004_40300	60.4%	87.3%
Day2_APS_005_28815	50.7%	60.5%
Day2_APS_006_3649	89.8%	91.9%
Day2_APS_007_20157	80.2%	82.3%
Day2_APS_008_3647	89.5%	93.8%
Day3_DE_002_4955	81.9%	95.3%
Day3_DE_003_29089	7.7%	43.4%
Day3_DE_004_40300	85.4%	94.6%
Day3_DE_005_28815	31.9%	54.3%
Day3_DE_006_3649	86.1%	93.8%
Day3_DE_007_20157	39.1%	54.0%
Day3_DE_008_3647	79.4%	87.0%

Table 3.2 . A. Gating scheme for purity estimates. B. Purity results from the second batch of differentiation.

Comparison of Time-Course Expression Profiles

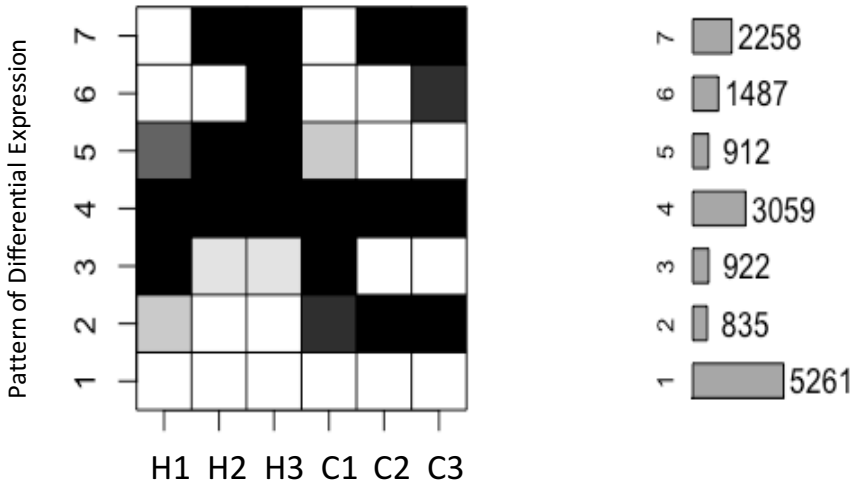
To classify genes by their expression profile over time, we jointly modeled all the data together using a Bayesian clustering approach to identify the most common temporal expression profiles or "core motifs"[98]. Over a third of genes (5261) are detected as unchanging through the time-course (cluster 1). Those whose expression levels change as differentiation progresses fall into six general patterns, four of which exhibit a similar profile in the human and chimpanzee time courses (clusters 3,4,7,6, Figure 3.4A.) For instance, genes that increase in expression only at definitive endoderm in humans (cluster 6) also do so in the chimpanzee. Several known regulators of primitive streak such as *Wnt3a* and *FoxH1* fall into cluster 3, with a high expression change at the transition between iPSC and primitive streak (Figure 3.4B). Notably, many genes fall into the same pattern cluster as the classic regulators of primitive streak, endoderm progenitor, and definitive endoderm, of which only a few dozen are well characterized. For instance, *ISOC1* and *TLE4* (Figure 3.4C) are not know to be involved in primitive streak specification, but the expression patterns of these genes assign them to cluster three with *Wnt3a* and *FoxH1*. Genes falling into cluster 7 share a pattern with genes that specify definitive endoderm and are significantly enriched for genes related to developmental disorders (Table 3.3)

We also note of particular interest are the categories 2 and 5, of roughly equal size, which correspond to genes whose expression changes significantly during the time-course only in chimpanzees and only in humans respectively. Genes in these clusters are candidates whose divergent regulation may give rise to ultimate phenotypic differences. Genes falling into cluster

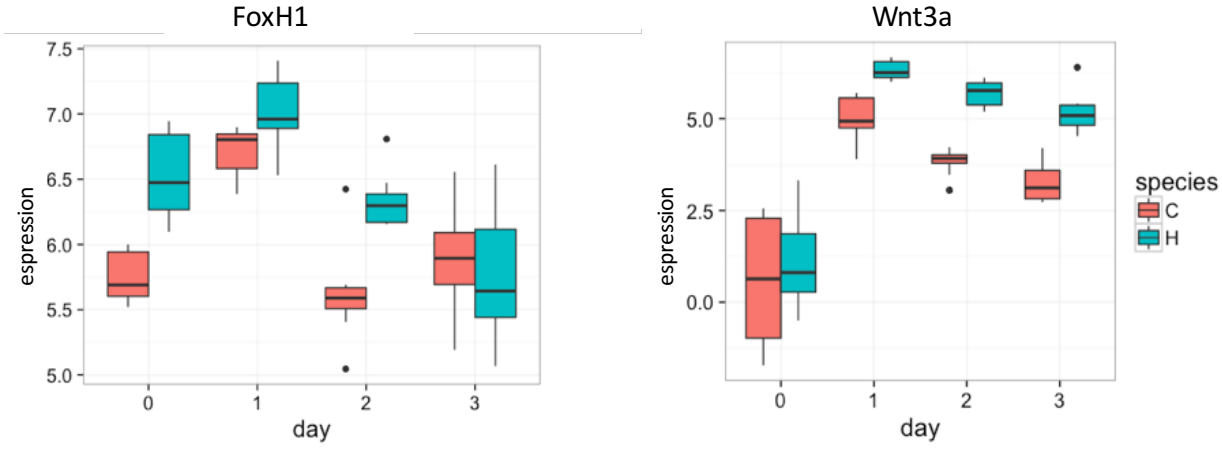
5, with changes in expression in humans but not chimpanzees as the time-course progresses include *SMOC2*, implicated in major dental malformations[99] and *ETV6* involved in salivary tumors[100], indicating differences in the regulation of these genes may be of interest for human health.

Figure 3.4: Bayesian Identification of Temporal Patterns of Global Gene Expression

A.



B.



C.

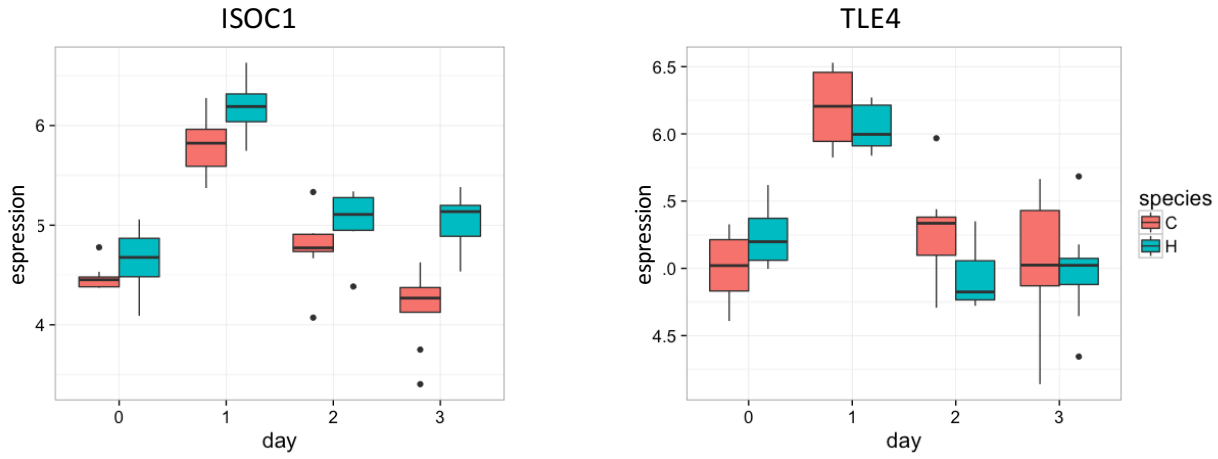


Figure 3.4 continued: A. Data was jointly modeled to identify clusters of genes whose expression changes similarly over time. Shading indicates the posterior probability of being differentially expressed compared to day 0. Comparisons are as follows: 1: Human Day 1 (H1) - H0, 2: H2 - H0, 3: H3 – H0, 4: Chimpanzee Day 1 (C1) - C0, 5: C2 - C0, 6: C3 - C0 B. Log_2 expression levels of FoxH1 and Wnt3a, primitive streak specifiers that fall into cluster three. C. Log_2 expression levels of ISOC1 and TLE4, which share an expression pattern with known primitive streak specifiers but whose role in primitive streak specification has not been studied.

		Phenotype Enrichment				
Feature ID	Name	Data Source	In Query	In Test Set	P-Value	Bayes
HP:0003128	Lactic acidosis	Human Phenotype Ontology	25	107	0.0000249	13.243
HP:0002151	Increased serum lactate	Human Phenotype Ontology	19	75	0.000265	10.156
HP:0001263	Global developmental delay	Human Phenotype Ontology	62	532	0.001	9.116
HP:0000007	Autosomal recessive inheritance	Human Phenotype Ontology	146	1717	0.012	6.971
HP:0000252	Microcephaly	Human Phenotype Ontology	51	455	0.027	5.522

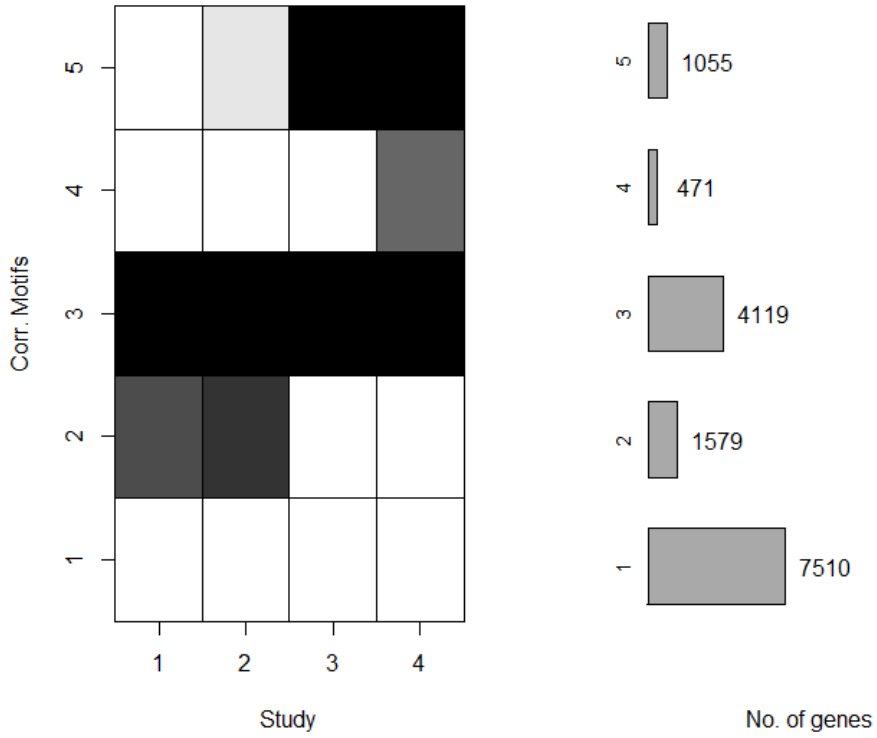
Table 3.3. Phenotype enrichment results from the Lynx database for genes falling into cluster 7.

To specifically examine the temporal dynamics of expression divergence, we next clustered genes by their probability of being differentially expressed between humans and chimpanzees

at each day (Figure 3.3a). Though the majority of genes are not differentially expressed between species at any time-point in this analysis, a substantial fraction are differentially expressed across the entire time-course (28%). The remainder are typically differentially expressed either early in the time course (11%, cluster 2), late in the time course (7%, cluster 5), or arising as DE only at the very end (3%, cluster 4). Genes whose differential regulation arises later in the time-course are of particular interest as potential agents of divergent physiology. We plot expression levels of a gene falling into this pattern, *APOE*, an apolipoprotein carrier of cholesterol which may have differential pathological consequences for the two primates.

Figure 3.5: Bayesian Idnetification of Differentially Expressed Genes Between Species Across the Time-course

A.



B.

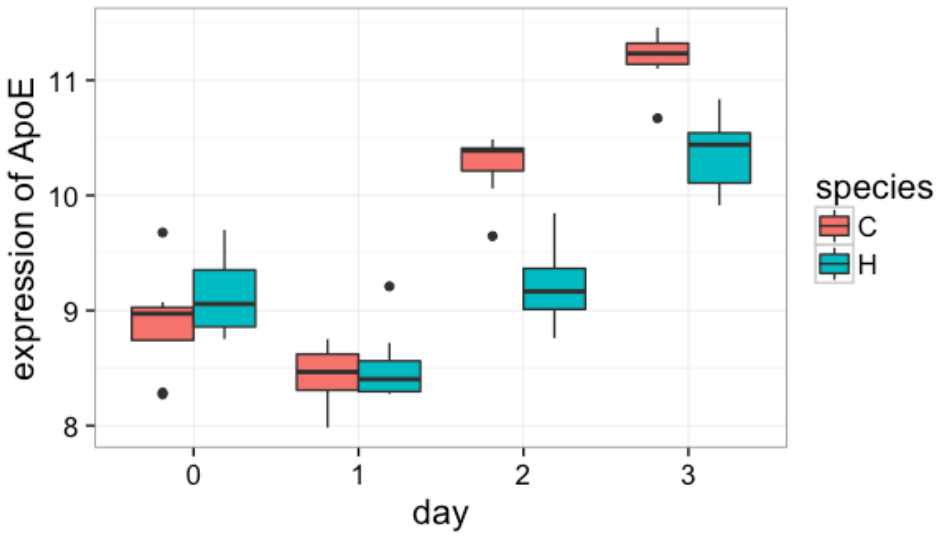
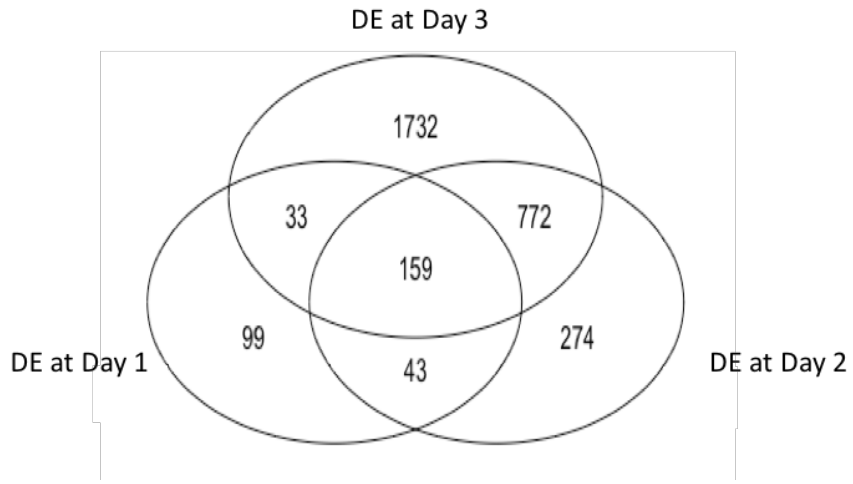


Figure 3.5 continued: A. Bayesian clustering of differentially expressed genes along the time-course using Cormotif. Shading indicates the posterior probability of being differentially expressed. B. Log_2 expression level of ApoE, a gene falling into pattern 5.

We further resolve the time-point at which expression levels diverge between the primates by identifying genes with a significant interaction between species and time-point at each day of the time-course, in comparison with expression level at day 0 (see methods for full details). The majority of genes with such an interaction exhibit a significant effect at day 3, and most do so uniquely (Figure 4A). However, the regulation of over 400 genes diverges between species during the middle days of the time-course while returning to equivalent expression by definitive endoderm (Figure 4B).

Figure 3.6: Genes with an interaction between Species with Time-point

A.



B.

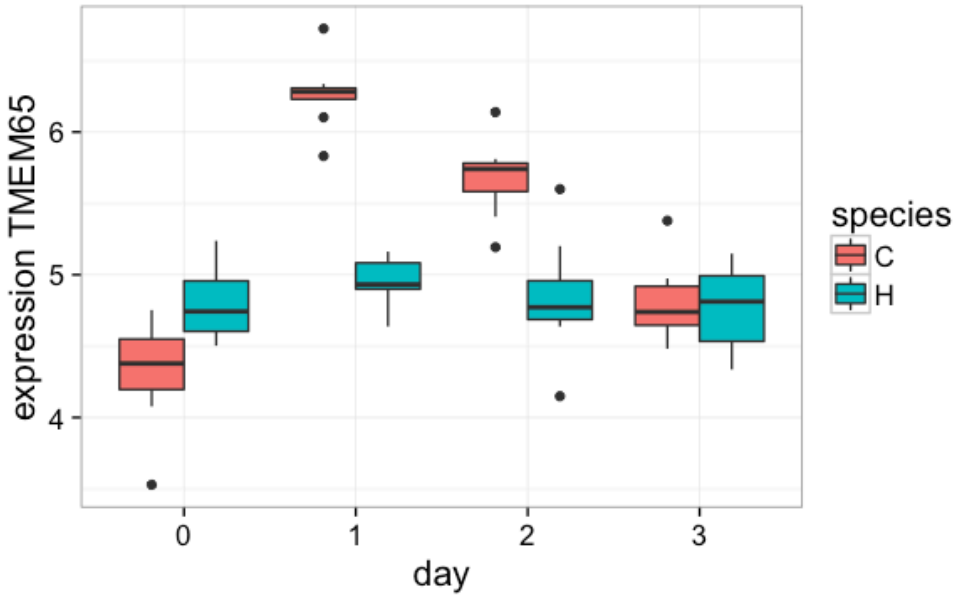


Figure 3.6 continued. A. Venn diagram of genes with a significant species-timepoint interaction effect by day (FDR<1%). B. Example of a gene with a significant species-timepoint interaction only at days 1 & 2.

To estimate the overall similarity of the expression state at each day, we quantified the correlation of the expression levels of all expressed genes at each day within and between the two species (Figure 3.7A). This assessment revealed that within a species, samples exhibit a significantly higher correlation at primitive streak than iPSC or definitive endoderm ($p < 0.05$ for all comparisons). Across species comparisons also indicate a more similar state between humans and chimpanzees at primitive streak than the preceding iPSCs ($p < 10^{-4}$), or the later stages of endoderm development ($p < 10^{-6}$ & $p < 10^{-13}$ respectively). The variance in global gene expression is also lowest at primitive streak than any other stage ($p < 10^{-6}$ for all time-point comparisons, Figure 3.7B).

With a reduction in variance, we expect to have more power to detect instances of differential expression with lower effect size and would therefore expect to see an increase in number of differentially expressed genes when this is the case. However, fewer genes are differentially expressed at primitive streak (Figure 3.7C), the stage with the lowest variance, than any other time-point and exceptionally few (276 genes) are uniquely differentially expressed at this time-point, suggesting that this stage is highly conserved (Figure 3.7 C and D).

Figure 3.7: Relative conservation of Developmental Stages

A

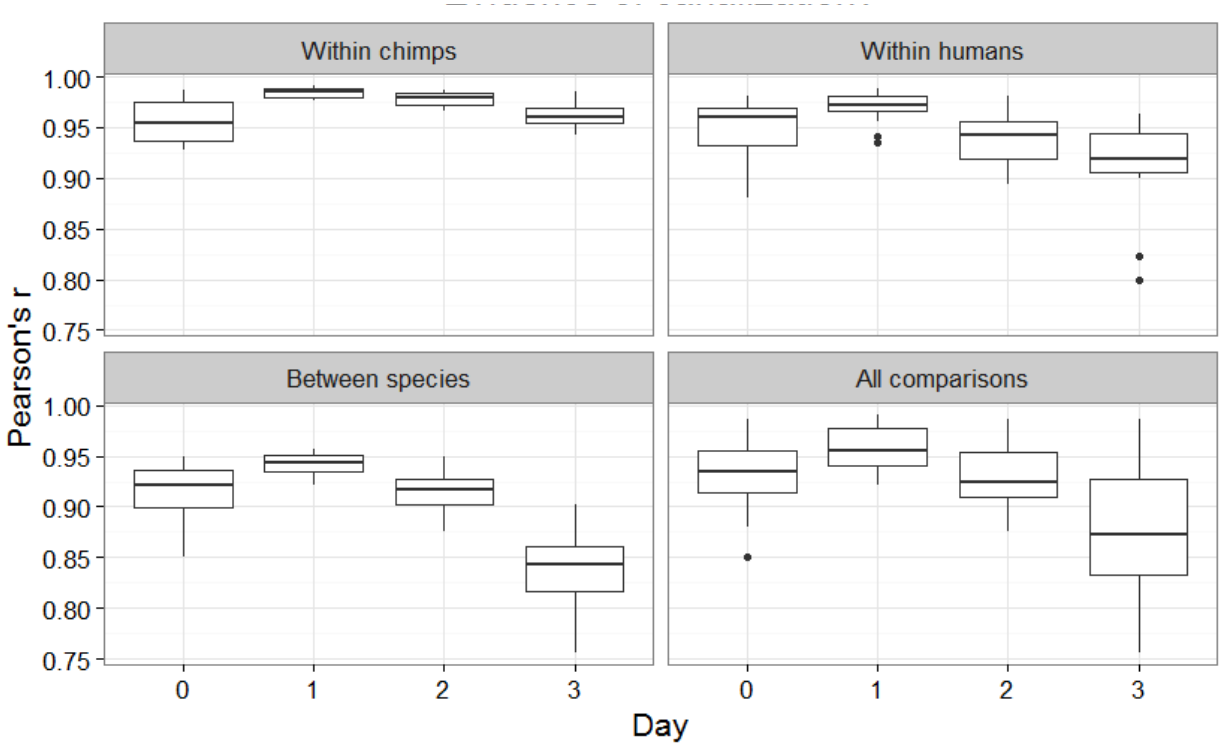


Figure 3.7 continued:

B.

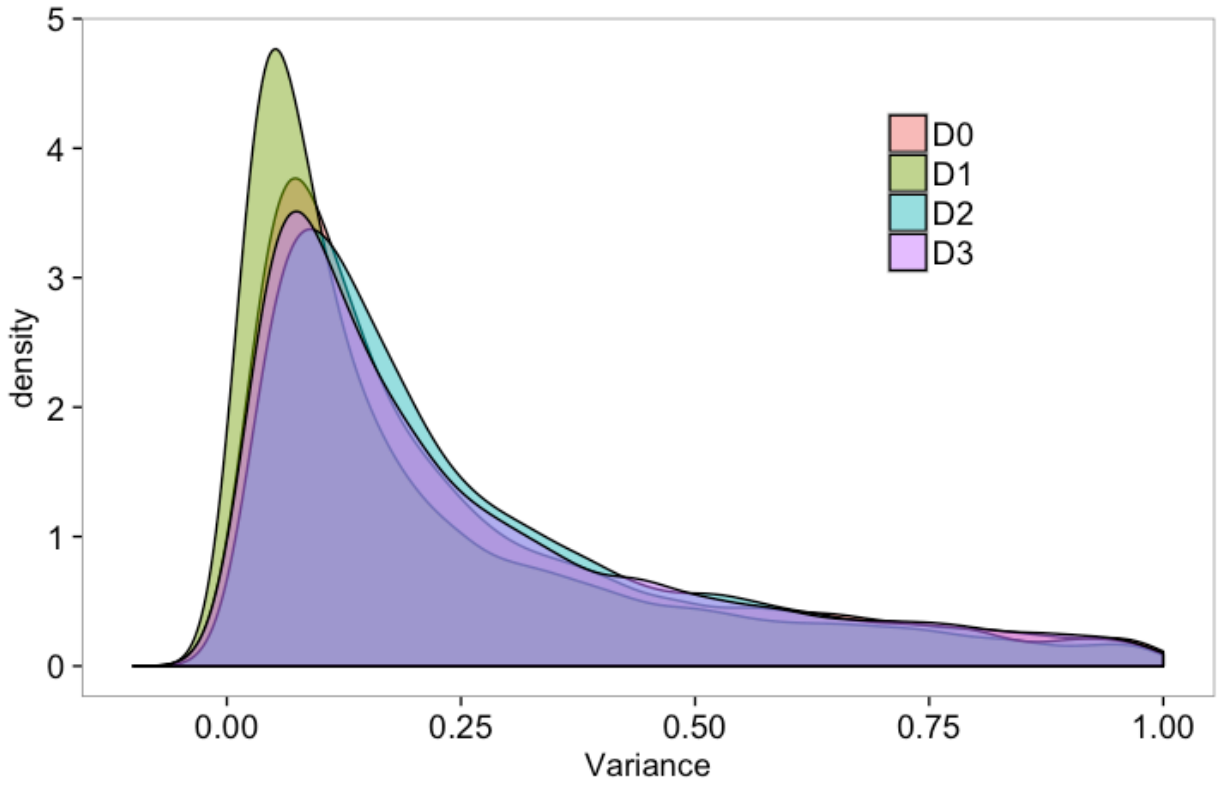
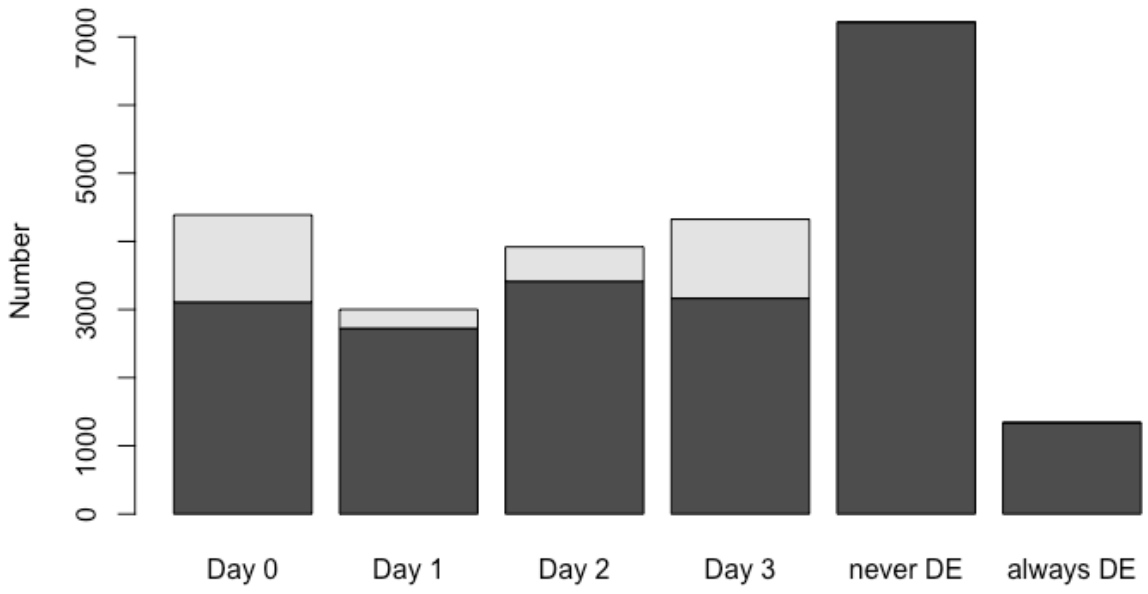


Figure 3.7 continued:

C.



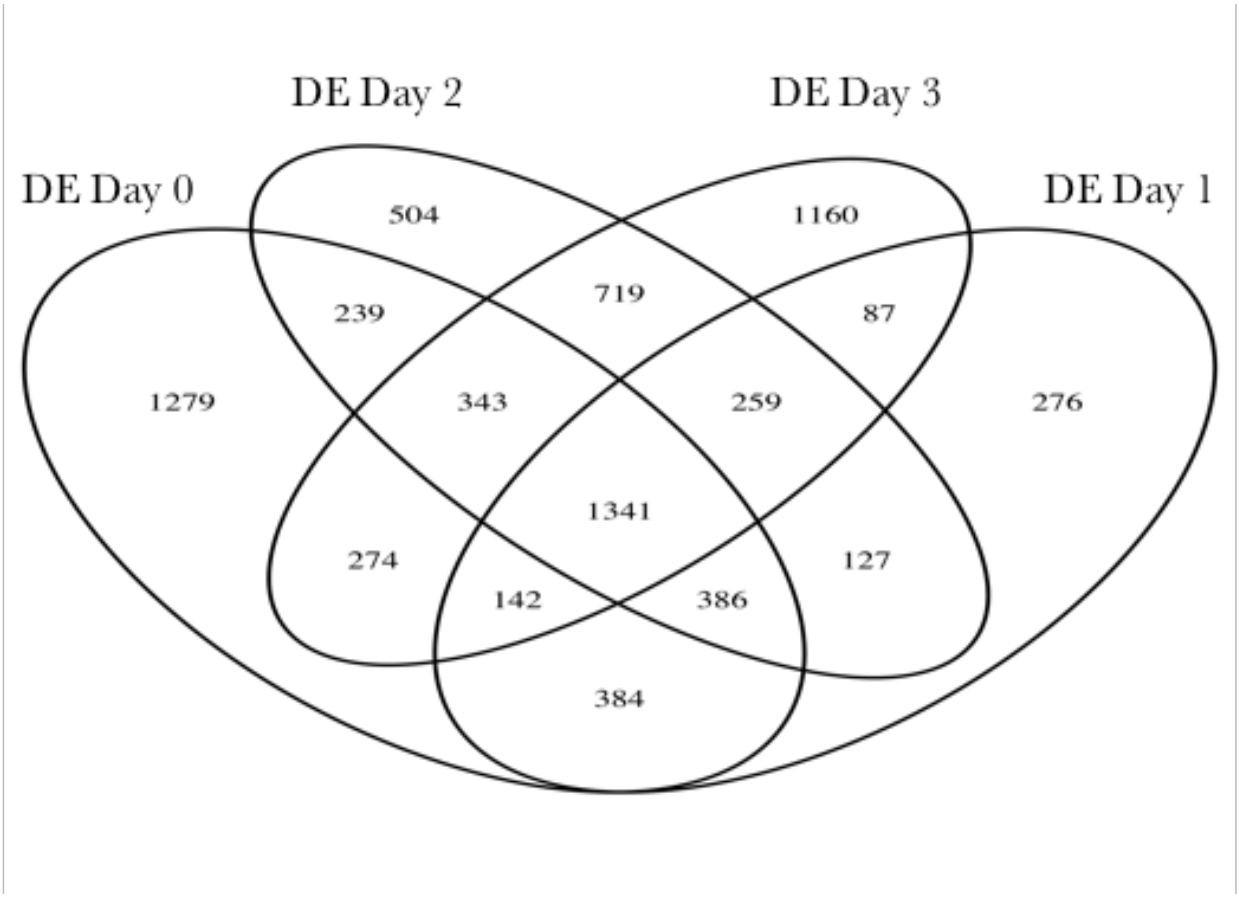


Figure 3.7 continued: A, Pairwise Pearson correlations between and within species at each time-point. B. Density plot of variance of $\log_2$ gene expression across all samples by day. Distribution tail cut off at x = 1.0. C. Number of genes differentially expressed between humans and chimpanzees at each time-point, never in the time-course, or at all time-points. White shading indicates the fraction of differentially expressed genes unique to the time-point. D. Venn diagram of all genes differentially expressed at each day.

3.4 Discussion

Here we report the first comparative study of early primate endoderm development using a panel of iPSCs from humans and chimpanzees. The endoderm germ layer is thought to be older than mesoderm or ectoderm and is the first layer to develop[101], Cells of endoderm origin ultimately form essential structures in the respiratory and digestive tracts including the liver, the pancreas, and the gall bladder, the lung, the thyroid, the bladder, the prostate, most of the pharynx and the lining of the auditory canals and the larynx. Because differences in these structures between humans and chimpanzees are thought to influence divergent primate characteristics, such as those related to nutrition and speech, a developmental understanding of the molecular divergence underlying these anatomical differences is of significant interest[102; 103]. On a broader level, we know very little about the relative molecular conservation of early developmental stages, especially in primates. Because many differences between primates are thought to be established early in development[104], this gap is a critical limitation in our understanding of human evolution. Thus, we studied global expression dynamics during the first four days of endoderm development to explore temporal divergence in gene regulation in the developing primate.

We find that the majority of genes whose precise regulation is known to be essential for the proper development of these early stages in the endoderm lineage, for instance *Sox2*, a regulator of pluripotency, *Brachyury*, a regulator of primitive streak, and *FoxA2*, a regulator of endoderm, are expressed at similar levels between humans and chimpanzees at the relevant day. This is expected as the roles of these genes in endoderm development were generally

identified in model organisms and are thus conserved in far more distantly related species than humans and chimpanzees. We also identify novel groups of genes whose expression pattern matches that of these known players and whose dynamic is also similar between the species. For instance, *ISOC1*, whose function is not well understood and *TLE4*, whose protein product interacts with PAX5[105], a member of the PAX gene family highly active during embryonic development[106] both fall into the cluster of genes whose expression pattern matches those of known primitive streak regulators. Likewise, genes falling into the cluster of genes whose expression changes only as cells commit to the endoderm lineage, in both primates, are enriched for genes involved in developmental disorders, indicating that they likely play a critical role in the achieval of proper development even though many of the roles have not been elucidated. Because of the conservation of their expression profiles, we suggest that these groups of genes may also play an important role in the developmental process and are particularly interesting targets for follow up as essential regulators of primate development. For such mechanistic follow-up studies, the iPSC system is a suitable alternative platform to model organisms as knock-down and gene editing experiments can be done more efficiently and at larger scales than in drosophila or zebrafish and certainly can complement work in model organisms.

Despite a high degree of overall expression conservation across species in early endoderm development, in fact nearly half of genes (49%) are never differentially expressed during the time-course, we highlight genes whose expression levels diverge early, late, or throughout the time-course. The differential regulation of these genes is a plausible mechanism for the ultimate divergence of adult physiology, especially those whose divergence is sustained as the

time-course progresses such as *APOE*, whose sequence differences between humans and chimpanzees have been well-studied and are predicted to have health consequences in risk for Alzheimer's and heart disease in humans[107].

On the other hand, we identify a fraction of genes that exhibit only a transient divergence early in our time-course. For instance, *TMEM65* falls into this category increasing in expression at primitive streak only in chimpanzees, while maintaining a constant low expression level in humans. *TMEM65* is known to be involved in cardiac development[108] but is not known to be active so early in the process. These may significantly influence the developing animal, for instance by adjusting the propensity for choosing certain paths at a bifurcation. Many such bifurcations and cell fate decisions occur during gastrulation. The ingression of the mesendoderm population through primitive streak, signifying the initiation of gastrulation, also represents a critical lineage bifurcation between mesoderm and endoderm[109]. Though the fewest genes are differentially expressed at this stage, those that are may have fundamental consequences, and are arguably more likely to result in significant anatomical differences than those DE later in the time-course.

Studies of gene regulation are often motivated by the understanding that such regulation varies greatly across tissues and contexts, such as time-point in development. However, interestingly, over 1300 genes are differentially expressed at every point of the time-course. EXAMPLE. These may be differentially expressed in a non-tissue specific pattern, and would therefore either be more likely to have larger effect on the physiology of the animal since more organ systems are

affected, or less likely if consequences of expression variation are so minor that they are not selected against.

It is of great interest to us what ultimate phenotypic consequences result from differential regulation of these subsets of genes. Future studies might explore tissue-specific expression differences of these categories in the adult animals, which is available from organs in every lineage for both humans and chimpanzees, and dissect their molecular roles by studying the functional consequences of differences in expression using knockout studies or gene-editing techniques *in vitro* in iPSCs or ESCs. Also, data-sets from many model organisms commonly used for developmental biology studies are readily available, and it is also feasible to compare the profiles of putative differentiation specifiers to other *in vivo* developmental datasets in model organisms.

Of more general interest, when taking a global assessment of our expression data, we observe the signature of canalization in our time-course, most pronounced at the primitive streak stage. Although the iPSC state is one of exceptionally low variance, as we report in previous studies[49], we find even lower variance in gene expression in primitive streak and the highest correlation estimates across the species. Subsequent stages fall into a more relaxed transcriptional state with higher variance and more differentially expressed genes. Such a profile is consistent with the activation of deeply conserved regulatory programs at the initial stages of endoderm development, followed by processes less affected by evolutionary constraint and therefore more accessible to adaptation.

Recent studies of gene expression during development have also addressed the question of relative conservation of developmental stages across species, several finding support for the existence of a phylotypic stage during mid-development where regulatory process are highly conserved[88; 91], and other distinctly not finding such evidence[90; 92]. Although our time-course is too early to compare early conservation levels to those of the alleged phylotypic stage, we propose that primitive streak may be an additional "bottleneck", at least in primate development. Perhaps development undergoes multiple phylotypic strictures, each representing a state of high evolutionary constraint, and, in its entirety, resembles something like a double hourglass. Presumably these highly restricted states exist to ensure precise regulation of crucial developmental process that must occur a particular way to result in a functional body plan and are not open to evolutionary interpretation. For instance, developmental junctures representing critical bifurcations in cell fate, to which great attention has recently been paid as targets in differentiation protocols from pluripotent stem cells [110; 111], may be examples of constrained processes. Primitive streak is such a stage, at which bifurcation into anterior and posterior PS is precarious, with fundamental consequences for the organism's body plan.

To our knowledge, this study is the first comparative assessment of global gene expression dynamics in early primate development. Using a highly controlled study design, we compared the chronology of overall molecular divergence of the developing endoderm in two closely related primate species and highlight primitive streak as a novel highly conserved state in early development. Using a Bayesian clustering approach, we also identify the predominant patterns of divergence between humans and chimpanzees during endoderm development and features

of genes that fall into these profiles. We suggest that an extension of this approach to other lineages would be highly rewarding and yield a new understanding of the divergent development of human-specific features of great interest such as those related to the brain, the skeletal structure, and the cardiovascular system.

3.2 Methods

iPSC panel

In this study, we include four chimpanzee iPSC lines (2 males, 2 females) from a previously described panel[49] and six human lines (3 males, 3 females) matched for cell type of origin, reprogramming method, culture condition and closely matched to passage number (median passage was within 1 passage across species and batches). All iPSC lines were previously satisfactorily evaluated for pluripotency measures, differentiation potential, lack of integrations and normal karyotypes. Original chimpanzee samples were obtained from the Yerkes Primate Center under protocol 006–12. Each chimpanzee was replicated in an independent differentiation, and two of the humans were replicated. Replicates were sex-balanced both within and across species. Feeder free cultures were initially maintained on Growth Factor Reduced Matrigel® using Essential 8 Medium™ (E8) as previously described. After 10 passages in E8, all cell lines were transitioned to iDEAL feeder free medium that was prepared in house as specified previously[112]. Cell culture was conducted at 37°C, 5% CO_2, and atmospheric O_2.

Endoderm Differentiation

To produce definitive endoderm and intermediate cell types, we followed a recently published three-day protocol that systematically identified and targeted pathways involved in cell fate decisions, at critical junctures in endoderm development[96] with minimal modification. At 12 hours prior to initiating differentiation, iPSC lines at 70-90% confluence were seeded at a density of 50,000 cells/cm^2. Day 1 (Primitive streak induction) media included 100 ng/mL Activin A, 50 nM PI-103 (PI3K inhibitor), 2 nM CHIR99021 (Wnt agonist). Day 2-3 media included

100 ng/mL Activin A and 250 nM LDN-193189 (BMP inhibitor). Days 1-3 were cultured in 50/50 IMDM/F12 basal media supplemented with 0.5 mg/mL human albumin, 0.7 μg/mL Insulin, 15 μg/mL holo-Transferrin and 1% v/v chemically defined lipid concentrate. Two rounds of differentiation were performed, resulting in replicates for most samples. Cell culture was conducted at 37°C, 5% CO_2, and atmospheric O_2.

Flow Cytometry for Purity Assessment

Cells were dissociated using an EDTA based cell release solution, centrifuged at 200 x g for 5 minutes at 4°C and washed with PBS. Subsequently, 0.5-1 million cells were fixed and permeabilized using the Foxp3 / Transcription Factor Staining Buffer Set from eBioscience. Cells were fixed at 4°C for 30 minutes before washing once using FACS buffer (autoMACS® Running Buffer, Miltenyi Biotech). 150,000 cells were transferred to BRAND lipoGrade 96 well immunostaining plates and centrifuged at 200 x g for 5 minutes at 4°C. Cells were rinsed in FACS buffer then resuspended in the staining solution. A single master mix containing 1X Permeabilization buffer (eBioscience), BD Horizon Brilliant Stain Buffer and antibodies was prepared and 30 μL of this mix was added to each well containing cells. In order to estimate purity for each day of the time course we utilized a mixture of six different directly labeled antibodies: Oct3/4 (BV421 labeled clone 3A2A20, Biolegend), SOX2 (PerCP-Cy5.5 labeled clone O30-678, BDbio), SOX17 (Alexa 488 labeled clone P7-969, BDbio), EOMES (PE-Cy7 labeled clone WD1928, eBioscience),CKIT (APC labeled clone 104D2, Biolegend), CXCR4 (BV605 labeled clone 12G5, Biolegend). All antibodies were used at the manufacturer recommended dilution except CKIT and CXCR4, used at 1/10 of the manufacturer specified concentration (15 ng of each

antibody in final volume of 30 µL per staining). Cells were stained for 1 hour at 4°C and subsequently washed 3x using a solution of BD Horizon Brilliant Stain Buffer containing 1X Permeabilization buffer, on the final wash cells were resuspended in 100 µL FACS buffer for acquisition on a BD LSR II flow cytometer. After data acquisition compensation scaling was determined in FlowJo using data from single stained compensation beads (Life Technologies) that were stained and collected in parallel. Live, intact, single cells were gated based on FSC and SSC channels as previously described[96]. Day 0 iPSC purity was estimated by dual positive OCT3/4 and SOX2[113] as well as negative staining for EOMES. Day 1 primitive streak purity was estimated primarily based on EOMES Positive staining[95; 114] but also negative staining for SOX17. Day 2 endoderm progenitor purity was quantified by positive staining for SOX17 expression[115] (CKIT could also be used as its level actually peaks at day 2 rather than day 3 with equivalent results) and negative staining for CXCR4. Finally, day 3 definitive endoderm purity was estimated by double staining for CKIT and CXCR4[116]. For all time points, cells were stained with the full complement of markers; initial gates were defined using fluorescence intensity levels of an iPSC line as a biological negative control for days 1, 2, and 3. For day 0 (iPSCs), a definitive endoderm time point was used to quantify the biological negative for OCT3/4 and SOX2 fluorescence intensity. All iPSC lines regardless of species were at comparable fluorescence intensity levels, so we choose a representative chimp and human line to use as our standard for defining and refining all gates. Fully resolving all time points simultaneously required us to define high and low staining gates which were determined using the time points for that markers maximum and minimum fluorescence intensity. All gates were refined using the same two representative chimpanzee and human lines as used for determining biological

negatives, resulting in one universal gating scheme that was applied to both species and all time points. A complete gating scheme for is outlined in Table 1A & S. Figure 1 , with the final purity results for the second round of differentiation in Table 1B.

Expression Quantification and Processing

We collected RNA from iPSCs (day 0) prior to adding day 1 media, and then every 24 hours during the differentiation time-course for a total of 4 time points representing all intermediate cell populations from iPSCs to definitive endoderm. RNA was extracted using the ZR-Duet DNA/RNA MiniPrep kit (Zymo) with the addition of an on column DNAse I treatment step prior to RNA elution. RNA concentration and quality was estimated using the Agilent 2100 Bioanalyzer.

50 base pair single-end RNA-seq libraries were created for all samples, multiplexed, and sequenced on the Illumina HiSeq 4000 at the Functional Genomics Core at University of Chicago. Read quality was evaluated using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Human reads were mapped to the hg19 genome and chimpanzee reads to panTro3 using the SubRead aligner (version 1.4.6)[117], allowing for up to two mismatches. Only exons with an annotated chimpanzee ortholog were included in analyses, a method originally used and described in Blekhman et. al.[118]. Gene expression levels were quantified using the *feature counts* function in the R package SubRead[119]. Only genes with an expression level of $\log_2$CPM > -2 in at least 10 samples of each species were included in subsequent analysis. Results are robust to other reasonable gene inclusion criteria such as requiring genes to be expressed in all

samples at all time-points as well as other normalization methods such as trimmed mean of M-values. A total of 14,737 genes past all inclusion criteria and were subject to analysis. One iPSC outlier was removed prior to all analyses.

Differential Expression and Time-course Modeling

Differential expression was estimated using a linear model based empirical Bayes method implemented in the R package 'limma'[120]. For this analysis, expression levels were normalized using the *voom* function. For pair-wise time-point comparisons, species and time-point were modeled as fixed effect terms, and individual within a species as a random effect term. Since technical factors (RIN score, batch, sex, extraction date, extraction batch, and mastermix) were not confounded with species and did not contribute significantly to the first five principle components of variation, no other factors were included as covariates. To estimate the effect of the interaction between species and time-point, gene expression levels were quantified relative to the mean level of the respective individual over all time-points. Genes were considered differentially expressed with FDR-adjusted p values <0.01.

To cluster genes by their temporal patterns in our data, we used CorMotif, a method that jointly models multiple expression data sets and, unlike other available methods in this class, allows for data-set specific differential expression patterns[98]. To identify patterns in expression changes over time, expression levels are compared to those of day 0. We classify genes into patterns with a posterior likelihood estimate of > 0.9 to be called differentially expressed between species or time-points and < 0.5 to be considered not differentially expressed.

Functional group enrichment was assessed using the web-based gene annotation database Lynx: http://lynx.ci.uchicago.edu using all expressed genes subjected to our filtering criteria as background. Enrichment was determined using an FDR cutoff of 5%

CHAPTER 4

Discussion

A growing appreciation for the plasticity of cell fates reached a peak in 2006 when terminally differentiated fibroblasts were reprogrammed to pluripotency[27]. In the intervening decade, the iPSC model has become utterly pervasive, finding countless applications in disease modeling, regenerative medicine, drug discovery, evolutionary biology, and functional genetics. Never has a method so quickly been integrated into diverse fields and ignited a collective anticipation of its potential to shape the future of human research. The model's sudden pervasiveness, however, has also triggered caution in many, and meticulous investigations into the relative suitability of the adult cell types commonly used for reprogramming. We discuss these concerns here and the contribution of our work toward a consensus about the faithfulness of iPSCs as a model of human variation.

Comparative biology, in contrast, is a field whose imagination has historically exceeded the capabilities of its methods. Coarse anatomical differences between man and non-human primates were classified decades ago, but an understanding of the genetic origin underlying their pronounced divergence has been slow to come. Even the sequencing of the chimpanzee genome completed in 2005 has not led to the anticipated clarity, hindered primarily by our erratic ability to predict physiological consequences of regulatory sequence variation. Developments in iPSC methodology now present an opportunity to resolve the cellular

processes through which non-coding genetic changes give rise to species' distinguishing phenotypes and to understand how they have been such a powerful force in shaping human evolutionary history. Here we discuss the results of a time-series comparative analysis of early human and chimpanzee development and future directions for this approach in comparative biology.

4.1 iPSCs as a Faithful Model of Human Variation

Perhaps as a result of their unusually rapid acceptance, the molecular integrity of iPSCs has been the subject of suspicion since the model was first developed. Reprogramming adult cells to pluripotency requires a reversal of the natural process of differentiation, famously portrayed as a ball rolling downhill[121]. The idea that we are sending cells back uphill to their ephemeral pluripotent state in the early moments after fertilization is a difficult one to accept, made even more difficult when considering how straight-forward the laboratory protocol is. Thus, motivated by the fear that iPSCs are somehow inadequate in their suddenly acquired pluripotency, extensive studies on the genome and epigenome of iPSCs now nearly constitute a field of their own.

Chief among concerns has been the possibility that the iPSC epigenome fails to resemble that of hESCs, and that instead, iPSCs retain a degree of epigenetic similarity to their adult cell precursors. Because hESCs isolated from a human blastocyst represent the natural state of pluripotency *in vivo*, the great attention paid to their deviations from and similarities to the iPSC genome is justified. As a whole, the two cell types are remarkably alike. While both demonstrate a rather high rate of chromosomal abnormalities and focal mutations, these

estimates are roughly the same for hESCs and hiPSCs and are in the range expected of cultured cells[122]. Likewise, methylation and histone modification patterns are nearly identical between the cell types, though consistently not so in a small subset of genes including imprinted genes and those on the X chromosome[123]. These sites remain an area of active exploration, though differences from ESCs appear to be mitigated by sustained iPSC passaging[124]. A further indication that these sites are not critical to the pluripotency of iPSCs is the successful generation of an entire mouse from iPSCs through tetraploid complementation, the ultimate demonstration of pluripotency[125-127].

The related idea that iPSCs harbor a 'memory' of their precursor adult cell is also a serious one. It is expected, even desirable, that iPSCs will be used as a source of mature cells in an entirely different lineage than the cells used to create them. Early studies reported a disturbingly large effect of cell type of origin on the ultimate gene expression state of iPSCs. However, all initial study designs confounded cell type of origin with donor genotype making it impossible to attribute variance to the appropriate factor[128; 129], or included only one individual and therefore no estimate of genetic effect at all[55; 130]. Only recently were studies designed to partition the variance between genotype and cell type of origin.

Aside from being more scientifically sound, the intentional inclusion of genetic variation in study design represented an important shift in perspective. Instead of searching for any evidence of epigenetic memory, the question became how significant this effect would be relative to that of donor genotype on molecular phenotypes, a much more meaningful concept with respect to interpreting results. The first report designed to compare the respective

influence of technical factors found that while some genes may retain the imprinting status of their precursor cell type, genetic variation had by far the strongest impact on expression[75]. A second properly controlled study supported this conclusion, finding less than 0.1% of tested sites retaining the methylation pattern of precursor cells, resulting in only a single differentially expressed gene[56]. As study design has improved, epigenetic "memory" appears to contribute little to the total expression variation, overshadowed by the effect of donor genotype.

The epigenetic overhaul that accompanies the reprogramming process and the low level of precursor cell memory in the resulting iPSCs inspired us to investigate the effects of reprogramming a cell type well known to be genetically dysregulated and therefore increasingly avoided. Lymphoblastoid cell lines were initially used as a convenient source of DNA for genetic studies such as the HapMap project, then experienced a golden age of applications in functional genomics, pharmacogenomics, and disease research because of their availability in large, diverse panels[24]. Thousands of studies have used them as a model system for human health applications such as predicting toxicity and therapeutic response[131]. Their popularity has declined in recent years, however, after a series of concerns were raised regarding their ability to faithfully represent their donor's innate gene regulatory system[25; 132]. Prolonged culture appears to select for cells with a growth advantage, thus instead of embodying the qualities of their diverse donors, LCLs converge on a common phenotype most permissive of immortal growth[25].

We find that the genetic contribution to expression heterogeneity, lost during LCL maintenance, is regained upon reprogramming to pluripotency. iPSCs made from LCLs are

nearly indistinguishable from those made from other cell types, consistently demonstrate loss of genomic EBV, and readily go on to produce tissues in all lineages. While essentially equivalent in their potential to convert into viable iPSCs, LCLs have a clear advantage over other commonly used starting sources. During their years of popularity, human LCLs were collected in large banks by hundreds of labs and consortiums. Hundreds of thousands of LCLs made from diverse patient and ethnic groups are available in panels such as those of the Coriell Cell Repositories[133], the International Histocompatibility Working Group[134], and many more biobanks in Europe[135], Australia[136], and Asia[137; 138]. As iPSCs are replacing LCLs as a model, and can be directly produced from existing lines, the utility of these panels for disease modeling and genetic studies is even greater than originally envisioned. Several institutions have already announced their intentions to reprogram their LCL panels into iPSCs such as the National Institute for Mental Health[139] (>100,000 samples) and the Framingham Heart Study investigators[140] (>4,000 samples). The economy of these ventures cannot be overstated. LCL biobanks are poised as an immediate resource for iPSC studies and bypass the slow and costly process of patient recruitment and sample collection, potentially accelerating any iPSC study by several years.

4.2 iPSCs to Study Human Evolution and Development

As more animals are being sequenced it is increasingly clear that the vast majority of protein-coding genes are highly similar between species with many being perfectly conserved between humans and our primate cousins. Humans do not have more genes. Our genes aren't longer or more complicated. As first proposed by King and Wilson over 40 years ago[44], the characteristics we define ourselves most by- our intelligence, anatomy, capacity for speech and culture- are in

large part the result of spatial and temporal differences in the regulation of the same genes that are shared across the animal kingdom.

A wide, lively literature of gene regulation in primate tissue samples has begun to unearth compelling stories about our evolution. The first global study of gene expression comparisons between humans and chimpanzees, for instance, found that expression evolved faster along the human lineage in the brain than in other tissues[141]. Further explorations have identified trends of positive selection on expression levels in the human brain[142] and found differences in gene expression patterns during aging in man compared to chimpanzee[143]. Metabolic pathways have also been highlighted as particular targets of selection along the human lineage[4]. These instances of adaptation through regulatory mechanisms are intriguing and have promoted an appreciation for the influence of non-coding variation on our evolutionary history. However, the methodology inevitably limits conclusions to static expression levels with an unknown environmental effect and precludes the direct experimentation required to study the operational principles of dynamic regulation. An additional difficulty is the critical scarcity of chimpanzee tissue samples. Very few organ biopsies have been collected for expression studies and, because of recent federal regulations to protect this endangered species, these samples are likely all that will ever be available. With respect to these limitations, iPSCs are a game changer. Panels of iPSCs from chimpanzees are a renewable source of the adult tissues that are in such limited supply, and can be directly experimented on and manipulated.

IPSC practices, sequencing technology, and developmental biology have all recently advanced to the point where it became possible, for the first time, to collect a high resolution gene expression time course profile of very early stages in primate development. This inceptive period from the pluripotent state of blastocyst cells to the appearance of distinct germ layers is a mysterious interval in any animal and especially little is known about this process in primates as modern ethics prohibits the relevant sample collection. It is entirely unknown if, when, and to what degree critical pathways become differentially activated during development in man compared to other primates. The time course design is powerful in this respect, enabling resolution of the origin of divergent molecular behavior, chronological patterns of differential expression, and the relative conservation of primitive cellular states.

Tissues arising from the endoderm germ layer account for characteristic differences between humans and other primates. All evidence indicates that diet underwent a great shift in the time since the human and chimpanzee lineages diverged, potentially accelerating adaptation. In fact, recent studies suggest that these changes influenced the selective pressure on genes involved in metabolism and nutrition[46]. From anatomical studies we know that the relative proportions of segments of the human gastrointestinal tract are far different than the chimpanzee with the proportion of man's small intestine being three times larger than the chimpanzee while that of the large intestine is less than half. Also of particular interest as evolutionary targets are structures related to speech of endoderm origin. The molecular developmental processes underlying differences in these structures and tissues is entirely unstudied.

We find that in the early developing endoderm, the expression levels of genes whose roles are known to be critical to early endoderm development from studies of model organisms are also generally conserved between human and chimpanzees. In fact, most genes are never significantly differentially expressed (DE) between the primates during the time-course. We identify genes in this category that also share an expression pattern with master regulators of endoderm development and are thus potential key regulators of primate endoderm development. Indeed, we find that these gene sets are enriched in genes whose disruption is associated with developmental disorders. Genes that are significantly DE are usually only so in a subset of the time-course which we divide into "early" and "late". Many genes begin as DE in iPSCs or arise as DE in definitive endoderm, however the intervening time-points are relatively well-conserved, exhibiting higher cross-species correlation and lower variance.

In this regard, our work engages another old but enduring theory, put forward in the 1940-50s by Conrad Waddington in his proposal that development is "canalized" and that this canalization is fundamentally linked to evolutionary processes[144]. Returning to the earlier metaphor where we imagine a cell in a developing organism is a ball rolling downhill, canalization is the idea that the ball rolls in predetermined tracks bordered by high edges, ensuring that cells can choose only from a finite number of end states. This setup results in an element of insensitivity to the environment as minor fluctuations do not produce an altogether new cell type, which is intuitively appealing considering the discrete nature of cell types in an adult animal.

Though many believe the phenomenon of canalization to be essential to evolutionary processes, the mechanism of its contribution is not obvious. It may be that canalization generally acts as a hindrance to evolutionary change, as a buffered system is less likely to be strongly affected by the introduction of mutations. However, if genetic variation can accumulate without phenotypic effect, it will be "stored" in a sense, and accessed only upon a strong shift in environment or mutational load. This concept of evolutionary capacitance has been invoked to explain events of speciation under stressful environmental circumstances whereby latent variation is exposed and, if advantageous, fixed through genetic assimilation. However, no theory regarding the nature of the link between canalization and adaptation has acquired enough consistent evidence to be generally accepted[145].

Our results indicate variation in the degree of canalization as the time-course progresses . Although the initial pluripotent state is one of extremely low variance, inter-species variation is reduced even further relative to within-species variation once endoderm development is initiated. As development proceeds, variance increases again both within and between species, suggesting a more relaxed transcriptional state. This chronological profile could be explained by initial activation of highly conserved regulatory networks, protected from evolutionary forces by some buffering mechanism and thereby deeply canalized, followed by transference to pathways more accessible to adaptation.

The connection between canalization of the molecular processes of development and evolutionary mechanisms is intriguing, but controversial. Today there are two main theories competing to explain the temporal profile of cross-species conservation of expression states during embryonic development. The early conservation model, an interpretation of von Baer's

third law[85], suggests that divergence in gene expression increases as development progresses, with the earliest stages being the most highly conserved. This theory is motivated by the idea that the most events occurring during the initial stages of development are the most fundamental, and that any variation at these stages would propagate to profoundly affect the adult animal. However, the more popular model is the hourglass model, which proposes that expression states are most highly conserved during mid-development at the "phylotypic stage" where the developing animals exhibit morphological similarity across species, and the earlier and later stages of development are subject to fewer evolutionary constraints[146]. The compelling possibility here is that the processes occurring during mid-development, organogenesis and the establishment of the body plan, are critically important and have such fundamental consequences for the adult animal, that the molecular events, such as activation of the *Hox* genes, must be precisely regulated and are therefore deeply canalized. In contrast, there are many ways that an animal can begin to develop. Proponents of this theory cite the multitude of modes of oogenesis and fertilization that initiate development to argue that early developmental stages are actually highly divergent across animals, and it is not until mid-development that cross-species resemblance emerges.

Over the past five years, studies of gene expression during embryonic development have investigated the relative conservation of developmental stages and have also attempted to characterize the temporal profile of canalization, re-awakening the controversy initiated by developmental biologists studying morphological differences across species. One of the first studies analyzed gene expression during development across six drosophila species found the highest correlation in expression at mid-development, consistent with the hourglass model and

a highly conserved phylotypic stage[88]. An analysis of gene age during zebrafish development found the oldest genes to be expressed during mid-development[92] . However, other studies have found no increase in gene age or cross-species expression correlations during mid-development[91], and in fact one study of development in animals across ten phyla found a decrease in expression correlation during mid-development[90]. The issue of the conservation of early developmental processes is far from resolved[147] and notably, has never been studied across primates.

We propose that during our time-course, we encounter another "bottleneck" of expression conservation at the primitive streak stage, before mid-development. Similar to the theory underlying the existence of the phylotypic stage, we argue that the processes occurring during primitive streak formation and ingression are critical to the proper establishment of the adult animal body plan, and are more complicated than the processes associated with earlier stages of development, where cells are maintaining pluripotency, not yet committing to lineages, and are essentially just dividing. On the other hand, the exact positioning of cells within the primitive streak determines ultimate cell fate, and, specifically, the movements of definitive endoderm induce normal body plan patterning[109]. Once the three germ layers are established, we propose that developmental constraints are less conserved and progenitor populations differ across species to give rise to the great variation in body plans. There is evidence for this pattern during early development in the analysis of gene age during zerbrafish development, which found an increase in the expression of older genes immediately prior to gastrulation followed by a peak in young transcripts, supporting the thesis that primitive streak may be an especially conserved stage.

Although our study does not fully support an early conservation model, in which embryonic stages become progressively divergent as development proceeds, it is consistent with an alternative interpretation of the model. Currently, the life cycle of an animal, at least as far as this debate is concerned, begins with egg formation and fertilization, the stages whose diversity of manifestations motivated the proposal of the hourglass model. However, these stages can also be viewed as one of the latest events of the life cycle occurring in the adult animal, the mother[147]. Under this interpretation of the early conservation model, global expression would be expected to diverge in the earliest stages of embryonic development, particularly those occurring before fetal gene expression is activated and maternal transcripts are still present, as they occur during the adult stage. In this case, the formation of primitive streak may very well be one of the first processes to occur in the embryo proper, and not as an extension of the life cycle of the mother. Our data is in accordance with this alternative interpretation, and we would be very interested to learn how it compares with data from other phyla, which currently have not been studied at the resolution required for discernment of distinct stages so early in the process of development.

4.3 Future and Limitations of the iPSC System for Studies of Gene Regulation

An important future extension of our work is to explore the role of cell by cell heterogeneity in endoderm development by studying single-cell gene expression in each of our populations. Recent studies of gene expression during development have investigated cellular heterogeneity, and consider it as a potential critical factor in cell fate determination. For

example, a study of single-cell gene expression during the first four days of mouse development revealed a progressive increase in variability in expression level of lineage specifying transcription factors from blastomeres in the 8 cell stage to cells in the inner cell mass. The higher variability in the timing of TF expression changes in the inner cell mass is thought to underlie fate decisions between primitive endoderm and epiblast commitment[148]. Similarly, a single cell gene expression study in humans pre-implantation embryos has proposed that heterogeneity in the downregulation of genes in human blastomeres contributes to lineage predispositions[149].

Studying the expression state of our cell populations at a single cell level is likely especially critical for the endoderm lineage. A recent study of the single cell gene expression of differentiation of embryonic stem cells down each of the three lineages, endoderm, mesoderm, and ectoderm, revealed that cells differentiating down the endoderm lineage exhibit the most asynchrony, indicating that there is likely substantial heterogeneity of degree of differentiation at any given time-point[150]. Similar to our results, this analysis detected a strong signal of distinct differentiation stages along the first three days of endoderm development, with only the final two days (three and four) overlapping in principal components analysis. Both their results and ours are consistent with passage through an intermediate multipotent “mesendoderm” population that is expected from the developmental biology literature and therefore increases confidence in the model. The interpretation of the overlap of the final time-points was that cell populations were beginning to stabilize at definitive endoderm by 72 hours and reached complete stabilization of the mature germ layer at 96 hours.

Spatial position in the embryo also contributes significantly to the cell-fate decision process, especially for cells that have ingressed through the primitive streak, and can be partially addressed by single-cell expression studies, even *in vitro*. A study of mesoderm differentiation in the developing mouse, for instance, used a method to place cells in pseudospace by classifying them by their expression of genes known to map across the anterior-posterior axis of the primitive streak (i.e. anterior primitive streak mesoderm cells interact with endoderm and are enriched for pathways involved in endoderm development)[151]. The technique is similar to methods used to place cells in pseudotime, and a combination of these approaches can partially address the limitation of the *in vitro* nature of the iPSC system, especially if these approaches are trained on in-vivo data from model organisms. Single-cell resolution of our comparative dataset would therefore be extremely advantageous for our study, in partitioning the differential expression into differences in timing of arrival at a given stage, differences in positioning trends, or, accounting for these factors, a difference in the cellular biology of the developmental stage itself.

An important limitation of our work is that species specific differences in developmental signaling are not attended to by modifications of the protocol between humans and chimpanzees. Any natural cell-driven developmental processes are likely overridden by our administered media conditions. Realistically however, current stem cell differentiation protocols are inevitably limited by an incomplete knowledge of development and likely operate with far lower sensitivity than would be required to capture species differences. Furthermore, in a canalized system, a rough approximation of the natural signal may be sufficient to achieve the same cell type, though perhaps this is overly optimistic. In any case, with this limitation in

mind, we emphasize the degree and quality of conservation throughout the time course instead of the importance of any particular pathway or gene.

We also stress that despite their many theoretical advantages, iPSCs are nevertheless an *in vitro* system which have fundamental drawbacks. The physiological environment of the body cannot be perfectly simulated in cell culture and therefore the measured cellular responses are not an exact reproduction of those occurring *in vivo*. While such an intrinsic limitation may never be completely overcome, it can be mitigated or avoided entirely by judicious experimental design. For instance, the effect of genotype has been detected *in vitro* by studying the differential response of isogenic cell lines, estimated independently from the overall state of the cell. The incorporation of pseudotime and pseudospace approaches in single-cell studies of development also mitigate the issue. The impact of this limitation will vary between applications, while many genetic studies can be confidently carried out, those using iPSCs for drug development and toxicity prediction will need to be more cautious. Although these are certainly considered among the disadvantages of *in vitro* work, the advantages far outweigh them, including a level of control over the system that cannot be achieved in in vivo designs and unparalleled accesses to time-points of development that will probably never be ethically accessible in humans and other primates. These considerations make in vitro work such as that presented here a valuable approach offering unique insights into human biology and evolution.

In this thesis, we present a quality assessment of induced pluripotent stem cells as a faithful model system for studies of human genetics and initial results from a comparative time course experiment of early primate development. There is every reason to believe that iPSCs will

advance the fields of comparative biology and human genetics, and equip them to expand in new directions. Our work is part of the beginning of what will be a powerful and productive approach in future genetics research.

References

1. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, NY) 337, 1190-1195.

2. Boyd, J.L.W., G. A. (2014). Evolution of Human Gene Expression Control. In. (eLS.

3. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325, 1246-1250.

4. Blekhman, R., Oshlack, A., Chabot, A.E., Smyth, G.K., and Gilad, Y. (2008). Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genet 4, e1000271.

5. Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440, 242-245.

6. Budczies, J., Weichert, W., Noske, A., Müller, B.M., Weller, C., Wittenberger, T., Hofmann, H.P., Dietel, M., Denkert, C., and Gekeler, V. (2011). Genome-wide gene expression profiling of formalin-fixed paraffin-embedded breast cancer core biopsies using microarrays. J Histochem Cytochem 59, 146-157.

7. Chiappini, F., Barrier, A., Saffroy, R., Domart, M.C., Dagues, N., Azoulay, D., Sebagh, M., Franc, B., Chevalier, S., Debuire, B., et al. (2006). Exploration of global gene expression in human liver steatosis by high-density oligonucleotide microarray. Lab Invest 86, 154-165.

8. Meugnier, E., Faraj, M., Rome, S., Beauregard, G., Michaut, A., Pelloux, V., Chiasson, J.L., Laville, M., Clement, K., Vidal, H., et al. (2007). Acute hyperglycemia induces a global downregulation of gene expression in adipose tissue and skeletal muscle of healthy subjects. Diabetes 56, 992-999.

9. Cooper-Knock, J., Kirby, J., Ferraiuolo, L., Heath, P.R., Rattray, M., and Shaw, P.J. (2012). Gene expression profiling in human neurodegenerative disease. Nat Rev Neurol 8, 518-530.

10. Sanz-Pamplona, R., Berenguer, A., Cordero, D., Riccadonna, S., Solé, X., Crous-Bou, M., Guinó, E., Sanjuan, X., Biondo, S., Soriano, A., et al. (2012). Clinical value of prognosis gene expression signatures in colorectal cancer: a systematic review. PLoS One 7, e48877.

11. Reis-Filho, J.S., and Pusztai, L. (2011). Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet 378, 1812-1823.

12. Oliver, P.L., Bitoun, E., and Davies, K.E. (2007). Comparative genetic analysis: the utility of mouse genetic systems for studying human monogenic disease. Mamm Genome 18, 412-424.

13. Wilson, J.M. (1996). Animal models of human disease for gene therapy. J Clin Invest 97, 1138-1141.

14. Wilke, M., Buijs-Offerman, R.M., Aarbiou, J., Colledge, W.H., Sheppard, D.N., Touqui, L., Bot, A., Jorna, H., de Jonge, H.R., and Scholte, B.J. (2011). Mouse models of cystic fibrosis: phenotypic analysis and research applications. J Cyst Fibros 10 Suppl 2, S152-171.

15. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768-772.

16. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., and Pritchard, J.K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12, R10.

17. Degner, J.F., Pai, A.a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482, 390-394.

18. Cheung, V.G., Spielman, R.S., Ewens, K.G., Weber, T.M., Morley, M., and Burdick, J.T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. Nature 437, 1365-1369.

19. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. PLoS Genet 10, e1004663.

20. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet 8, e1002639.

21. Hu, V.W., Frank, B.C., Heine, S., Lee, N.H., and Quackenbush, J. (2006). Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. BMC Genomics 7, 118.

22. Washizuka, S., Iwamoto, K., Kakiuchi, C., Bundo, M., and Kato, T. (2009). Expression of mitochondrial complex I subunit gene NDUFV2 in the lymphoblastoid cells derived from patients with bipolar disorder and schizophrenia. Neurosci Res 63, 199-204.

23. Consortium, I.H. (2003). The International HapMap Project. Nature 426, 789-796.

24. Sie, L., Loong, S., and Tan, E.K. (2009). Utility of lymphoblastoid cell lines. J Neurosci Res 87, 1953-1959.

25. Calışkan, M., Pritchard, J.K., Ober, C., and Gilad, Y. (2014). The effect of freeze-thaw cycles on gene expression levels in lymphoblastoid cell lines. PLoS One 9, e107166.

26. Caliskan, M., Cusanovich, D.A., Ober, C., and Gilad, Y. (2011). The effects of EBV transformation on gene expression levels and methylation profiles. Hum Mol Genet 20, 1643-1652.

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

28. Soares, F.A., Sheldon, M., Rao, M., Mummery, C., and Vallier, L. (2014). International coordination of large-scale human induced pluripotent stem cell initiatives: Wellcome Trust and ISSCR workshops white paper. Stem Cell Reports 3, 931-939.

29. Gallego Romero, I., Pavlovic, B.J., Hernando-Herraez, I., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., Mitrano, A., Chavarria, C.I., Ben-Nun, I.F., et al. (2014). Generation of a Panel of Induced Pluripotent Stem Cells From Chimpanzees: a Resource for Comparative Functional Genomics. bioRxiv.

30. Consortium, H.i. (2012). Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. Cell Stem Cell 11, 264-278.

31. Juopperi, T.A., Kim, W.R., Chiang, C.H., Yu, H., Margolis, R.L., Ross, C.A., Ming, G.L., and Song, H. (2012). Astrocytes generated from patient induced pluripotent stem cells recapitulate features of Huntington's disease patient cells. Mol Brain 5, 17.

32. Quarto, N., Leonard, B., Li, S., Marchand, M., Anderson, E., Behr, B., Francke, U., Reijo-Pera, R., Chiao, E., and Longaker, M.T. (2012). Skeletogenic phenotype of human Marfan embryonic stem cells faithfully phenocopied by patient-specific induced-pluripotent stem cells. Proc Natl Acad Sci U S A 109, 215-220.

33. Yagi, T., Ito, D., Okada, Y., Akamatsu, W., Nihei, Y., Yoshizaki, T., Yamanaka, S., Okano, H., and Suzuki, N. (2011). Modeling familial Alzheimer's disease with induced pluripotent stem cells. Hum Mol Genet 20, 4530-4539.

34. Kondo, T., Asai, M., Tsukita, K., Kutoku, Y., Ohsawa, Y., Sunada, Y., Imamura, K., Egawa, N., Yahata, N., Okita, K., et al. (2013). Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Aβ and differential drug responsiveness. Cell Stem Cell 12, 487-496.

35. Kiskinis, E., Sandoe, J., Williams, L.A., Boulting, G.L., Moccia, R., Wainger, B.J., Han, S., Peng, T., Thams, S., Mikkilineni, S., et al. (2014). Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. Cell Stem Cell 14, 781-795.

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

37. DeRosa, B.A., Van Baaren, J.M., Dubey, G.K., Lee, J.M., Cuccaro, M.L., Vance, J.M., Pericak-Vance, M.A., and Dykxhoorn, D.M. (2012). Derivation of autism spectrum disorder-specific induced pluripotent stem cells from peripheral blood mononuclear cells. Neurosci Lett 516, 9-14.

38. Israel, M.A., Yuan, S.H., Bardy, C., Reyna, S.M., Mu, Y., Herrera, C., Hefferan, M.P., Van Gorp, S., Nazor, K.L., Boscolo, F.S., et al. (2012). Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. Nature 482, 216-220.

39. Hua, H., Shang, L., Martinez, H., Freeby, M., Gallagher, M.P., Ludwig, T., Deng, L., Greenberg, E., Leduc, C., Chung, W.K., et al. (2013). iPSC-derived β cells model diabetes due to glucokinase deficiency. J Clin Invest 123, 3146-3153.

40. Burridge, P.W., Li, Y.F., Matsa, E., Wu, H., Ong, S.G., Sharma, A., Holmström, A., Chang, A.C., Coronado, M.J., Ebert, A.D., 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.

41. Hertz, D.L., Owzar, K., Lessans, S., Wing, C., Jiang, C., Kelly, W.K., Patel, J.N., Halabi, S., Furukawa, Y., Wheeler, H.E., et al. (2016). Pharmacogenetic Discovery in CALGB (Alliance) 90401 and Mechanistic Validation of a VAC14 Polymorphism That Increases Risk of Docetaxel-Induced Neuropathy. Clin Cancer Res.

42. Haygood, R., Babbitt, C.C., Fedrigo, O., and Wray, G.A. (2010). Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci U S A 107, 7853-7857.

43. Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8, 206-216.

44. King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science 188, 107-116.

45. Shibata, Y., Sheffield, N.C., Fedrigo, O., Babbitt, C.C., Wortham, M., Tewari, A.K., London, D., Song, L., Lee, B.K., Iyer, V.R., et al. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. PLoS Genet 8, e1002789.

46. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D., and Wray, G.A. (2007). Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet 39, 1140-1144.

47. Torgerson, D.G., Boyko, A.R., Hernandez, R.D., Indap, A., Hu, X., White, T.J., Sninsky, J.J., Cargill, M., Adams, M.D., Bustamante, C.D., et al. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet 5, e1000592.

48. Arbiza, L., Gronau, I., Aksoy, B.A., Hubisz, M.J., Gulko, B., Keinan, A., and Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45, 723-729.

49. Gallego Romero, I., Pavlovic, B.J., Hernando-Herraez, I., Zhou, X., Ward, M.C., Banovich, N.E., Kagan, C.L., Burnett, J.E., Huang, C.H., Mitrano, A., et al. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife 4, e07103.

50. Marchetto, M.C.N., Muotri, A.R., and Gage, F.H. (2013). Proposing a Model for Studying Primate

Development Using Induced Pluripotent

Stem Cells. In Programmed Cells from Basic Neuroscience

to Therapy. (Research and Perspectives in Neurosciences.

51. Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell 163, 68-83.

52. Otani, T., Marchetto, M.C., Gage, F.H., Simons, B.D., and Livesey, F.J. (2016). 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. Cell Stem Cell 18, 467-480.

53. Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134, 25-36.

54. Ghosh, Z., Wilson, K.D., Wu, Y., Hu, S., Quertermous, T., and Wu, J.C. (2010). Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. PLoS One 5, e8975.

55. Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature 467, 285-290.

56. Kagan, C.L., Banovich, N.E., Pavlovic, B.J., Patterson, K., Gallego Romero, I., Pritchard, J.K., and Gilad, Y. (2015). Genetic Variation, Not Cell Type of Origin, Underlies Regulatory Differences in iPSCs. bioRxiv.

57. Huang, R.S., Duan, S., Kistner, E.O., Hartford, C.M., and Dolan, M.E. (2008). Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. Molecular cancer therapeutics 7, 3038-3046.

58. Wen, Y., Gamazon, E.R., Bleibel, W.K., Wing, C., Mi, S., McIlwee, B.E., Delaney, S.M., Duan, S., Im, H.K., and Dolan, M.E. (2012). An eQTL-based method identifies CTTN and ZMAT3 as pemetrexed susceptibility markers. Hum Mol Genet 21, 1470-1480.

59. Ziliak, D., O'Donnell, P.H., Im, H.K., Gamazon, E.R., Chen, P., Delaney, S., Shukla, S., Das, S., Cox, N.J., Vokes, E.E., et al. (2011). Germline polymorphisms discovered via a cell-based, genome-wide approach predict platinum response in head and neck cancers. Transl Res 157, 265-272.

60. Moyer, A.M., Fridley, B.L., Jenkins, G.D., Batzler, A.J., Pelleymounter, L.L., Kalari, K.R., Ji, Y., Chai, Y., Nordgren, K.K.S., and Weinshilboum, R.M. (2011). Acetaminophen-NAPQI hepatotoxicity: a cell line model system genome-wide association study. Toxicological sciences : an official journal of the Society of Toxicology 120, 33-41.

61. Wheeler, H.E., and Dolan, M.E. (2012). Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. Pharmacogenomics 13, 55-70.

62. Peters, E.J., Kraja, A.T., Lin, S.J., Yen-Revollo, J.L., Marsh, S., Province, M.A., and McLeod, H.L. (2009). Association of thymidylate synthase variants with 5-fluorouracil cytotoxicity. Pharmacogenet Genomics 19, 399-401.

63. Gamazon, E.R., Huang, R.S., Cox, N.J., and Dolan, M.E. (2010). Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci U S A 107, 9287-9292.

64. Hartford, C.M., Duan, S., Delaney, S.M., Mi, S., Kistner, E.O., Lamba, J.K., Huang, R.S., and Dolan, M.E. (2009). Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. Blood 113, 2145-2153.

65. Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C., et al. (2008). Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. PLoS Genet 4, e1000287.

66. Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T., and Todd, J.A. (2008). Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. PLoS One 3, e2966.

67. Stark, A.L., Zhang, W., Mi, S., Duan, S., O'Donnell, P.H., Huang, R.S., and Dolan, M.E. (2010). Heritable and non-genetic factors as variables of pharmacologic phenotypes in lymphoblastoid cell lines. Pharmacogenomics J 10, 505-512.

68. Hannula, K., Lipsanen-Nyman, M., Scherer, S.W., Holmberg, C., Höglund, P., and Kere, J. (2001). Maternal and paternal chromosomes 7 show differential methylation of many genes in lymphoblast DNA. Genomics 73, 1-9.

69. Carter, K.L., Cahir-McFarland, E., and Kieff, E. (2002). Epstein-barr virus-induced changes in B-lymphocyte gene expression. J Virol 76, 10427-10436.

70. Min, J.L., Barrett, A., Watts, T., Pettersson, F.H., Lockstone, H.E., Lindgren, C.M., Taylor, J.M., Allen, M., Zondervan, K.T., and McCarthy, M.I. (2010). Variability of gene expression profiles in human blood and lymphoblastoid cell lines. BMC Genomics 11, 96.

71. Powell, J.E., Henders, A.K., McRae, A.F., Wright, M.J., Martin, N.G., Dermitzakis, E.T., Montgomery, G.W., and Visscher, P.M. (2012). Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. Genome Res 22, 456-466.

72. Mills, J.A., Wang, K., Paluru, P., Ying, L., Lu, L., Galvão, A.M., Xu, D., Yao, Y., Sullivan, S.K., Sullivan, L.M., et al. (2013). Clonal genetic and hematopoietic heterogeneity among human-induced pluripotent stem cell lines. Blood 122, 2047-2051.

73. Boulting, G.L., Kiskinis, E., Croft, G.F., Amoroso, M.W., Oakley, D.H., Wainger, B.J., Williams, D.J., Kahler, D.J., Yamaki, M., Davidow, L., et al. (2011). A functionally characterized test set of human induced pluripotent stem cells. Nat Biotechnol 29, 279-286.

74. Kajiwara, M., Aoi, T., Okita, K., Takahashi, R., Inoue, H., and Takayama, N. (2012). Correction for Kajiwara et al., Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proceedings of the National Academy of Sciences 109, 14716-14716.

75. Rouhani, F., Kumasaka, N., de Brito, M.C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS Genet 10, e1004432.

76. Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K., et al. (2011). A more efficient method to generate integration-free human iPS cells. Nat Methods 8, 409-412.

77. Choi, S.M., Liu, H., Chaudhari, P., Kim, Y., Cheng, L., Feng, J., Sharkis, S., Ye, Z., and Jang, Y.Y. (2011). Reprogramming of EBV-immortalized B-lymphocyte cell lines into induced pluripotent stem cells. Blood 118, 1801-1805.

78. Rajesh, D., Dickerson, S.J., Yu, J., Brown, M.E., Thomson, J.A., and Seay, N.J. (2011). Human lymphoblastoid B-cell lines reprogrammed to EBV-free induced pluripotent stem cells. Blood 118, 1797-1800.

79. Müller, F.J., Schuldt, B.M., Williams, R., Mason, D., Altun, G., Papapetrou, E.P., Danner, S., Goldmann, J.E., Herbst, A., Schmidt, N.O., et al. (2011). A bioinformatic assay for pluripotency in human cells. Nat Methods 8, 315-317.

80. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118-127.

81. Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3, Article3.

82. Sulakhe, D., Balasubramanian, S., Xie, B., Feng, B., Taylor, A., Wang, S., Berrocal, E., Dave, U., Xu, J., Börnigen, D., et al. (2014). Lynx: a database and knowledge extraction engine for integrative medicine. Nucleic Acids Res 42, D1007-1012.

83. Garfield, D.A., Runcie, D.E., Babbitt, C.C., Haygood, R., Nielsen, W.J., and Wray, G.A. (2013). The impact of gene expression variation on the robustness and evolvability of a developmental gene regulatory network. PLoS Biol 11, e1001696.

84. Roux, J., and Robinson-Rechavi, M. (2008). Developmental constraints on vertebrate genome evolution. PLoS Genet 4, e1000311.

85. Abzhanov, A. (2013). von Baer's law for the ages: lost and found principles of developmental evolution. Trends Genet 29, 712-722.

86. Sander, K., and Schmidt-Ott, U. (2004). Evo-devo aspects of classical and molecular data in a historical perspective. J Exp Zool B Mol Dev Evol 302, 69-91.

87. Richardson, M.K., and Keuck, G. (2002). Haeckel's ABC of evolution and development. Biol Rev Camb Philos Soc 77, 495-528.

88. Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. Nature 468, 811-814.

89. Hazkani-Covo, E., Wool, D., and Graur, D. (2005). In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. J Exp Zool B Mol Dev Evol 304, 150-158.

90. Levin, M., Anavy, L., Cole, A.G., Winter, E., Mostov, N., Khair, S., Senderovich, N., Kovalev, E., Silver, D.H., Feder, M., et al. (2016). The mid-developmental transition and the evolution of animal body plans. Nature 531, 637-641.

91. Piasecka, B., Lichocki, P., Moretti, S., Bergmann, S., and Robinson-Rechavi, M. (2013). The hourglass and the early conservation models--co-existing patterns of developmental constraints in vertebrates. PLoS Genet 9, e1003476.

92. Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468, 815-818.

93. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol 4, e72.

94. Olson, M.V., and Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 4, 20-28.

95. Teo, A.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes Dev 25, 238-250.

96. Loh, K.M., Ang, L.T., Zhang, J., Kumar, V., Ang, J., Auyeong, J.Q., Lee, K.L., Choo, S.H., Lim, C.Y., Nichane, M., et al. (2014). Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. Cell Stem Cell 14, 237-252.

97. D'Amour, K.A., Agulnick, A.D., Eliazer, S., Kelly, O.G., Kroon, E., and Baetge, E.E. (2005). Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat Biotechnol 23, 1534-1541.

98. Wei, Y., Tenzen, T., and Ji, H. (2015). Joint analysis of differential gene expression in multiple studies using correlation motifs. Biostatistics 16, 31-46.

99. Bloch-Zupan, A., Jamet, X., Etard, C., Laugel, V., Muller, J., Geoffroy, V., Strauss, J.P., Pelletier, V., Marion, V., Poch, O., et al. (2011). Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. Am J Hum Genet 89, 773-781.

100. Chiosea, S.I., Griffith, C., Assaad, A., and Seethala, R.R. (2012). Clinicopathological characterization of mammary analogue secretory carcinoma of salivary glands. Histopathology 61, 387-394.

101. Hashimshony, T., Feder, M., Levin, M., Hall, B.K., and Yanai, I. (2015). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature 519, 219-222.

102. Walker, A. (1981). Diet and teeth. Dietary hypotheses and human evolution. Philos Trans R Soc Lond B Biol Sci 292, 57-64.

103. Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3, e170.

104. Prabhakar, S., Visel, A., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Morrison, H., Fitzpatrick, D.R., Afzal, V., et al. (2008). Human-specific gain of function in a developmental enhancer. Science 321, 1346-1350.

105. Eberhard, D., Jiménez, G., Heavey, B., and Busslinger, M. (2000). Transcriptional repression by Pax5 (BSAP) through interaction with corepressors of the Groucho family. EMBO J 19, 2292-2303.

106. Gérard, M., Abitbol, M., Delezoide, A.L., Dufier, J.L., Mallet, J., and Vekemans, M. (1995). PAX-genes expression during human embryonic development, a preliminary report. C R Acad Sci III 318, 57-66.

107. McIntosh, A.M., Bennett, C., Dickson, D., Anestis, S.F., Watts, D.P., Webster, T.H., Fontenot, M.B., and Bradley, B.J. (2012). The apolipoprotein E (APOE) gene appears functionally monomorphic in chimpanzees (Pan troglodytes). PLoS One 7, e47760.

108. Sharma, P., Abbasi, C., Lazic, S., Teng, A.C., Wang, D., Dubois, N., Ignatchenko, V., Wong, V., Liu, J., Araki, T., et al. (2015). Evolutionarily conserved intercalated disc protein Tmem65 regulates cardiac conduction and connexin 43 function. Nat Commun 6, 8391.

109. Tam, P.P., and Loebel, D.A. (2007). Gene function in mouse embryogenesis: get set for gastrulation. Nat Rev Genet 8, 368-381.

110. Graf, T., and Enver, T. (2009). Forcing cells to change lineages. Nature 462, 587-594.

111. Loh, K.M., and Lim, B. (2011). A precarious balance: pluripotency factors as lineage specifiers. Cell Stem Cell 8, 363-369.

112. Marinho, P.A., Chailangkarn, T., and Muotri, A.R. (2015). Systematic optimization of human pluripotent stem cells media using Design of Experiments. Sci Rep 5, 9834.

113. Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A.A., et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625-635.

114. Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development 142, 2121-2135.

115. Viotti, M., Nowotschin, S., and Hadjantonakis, A.K. (2014). SOX17 links gut endoderm morphogenesis and germ layer segregation. Nat Cell Biol 16, 1146-1156.

116. Cheng, X., Ying, L., Lu, L., Galvão, A.M., Mills, J.a., Lin, H.C., Kotton, D.N., Shen, S.S., Nostro, M.C., Choi, J.K., et al. (2012). Self-renewing endodermal progenitor lines generated from human pluripotent stem cells. Cell stem cell 10, 371-384.

117. Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res 41, e108.

118. Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M., and Gilad, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. Genome Res 20, 180-189.

119. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923-930.

120. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15, R29.

121. Ferrell, J.E. (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. Curr Biol 22, R458-466.

122. Lund, R.J., Närvä, E., and Lahesmaa, R. (2012). Genetic and epigenetic stability of human pluripotent stem cells. Nat Rev Genet 13, 732-744.

123. Nazor, K.L., Altun, G., Lynch, C., Tran, H., Harness, J.V., Slavin, I., Garitaonandia, I., Müller, F.J., Wang, Y.C., Boscolo, F.S., et al. (2012). Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. Cell Stem Cell 10, 620-634.

124. Nishino, K., Toyoda, M., Yamazaki-Inoue, M., Fukawatase, Y., Chikazawa, E., Sakaguchi, H., Akutsu, H., and Umezawa, A. (2011). DNA methylation dynamics in human induced pluripotent stem cells over time. PLoS Genet 7, e1002085.

125. Boland, M.J., Hazen, J.L., Nazor, K.L., Rodriguez, A.R., Gifford, W., Martin, G., Kupriyanov, S., and Baldwin, K.K. (2009). Adult mice generated from induced pluripotent stem cells. Nature 461, 91-94.

126. Kang, L., Wang, J., Zhang, Y., Kou, Z., and Gao, S. (2009). iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. Cell Stem Cell 5, 135-138.

127. Zhao, X.Y., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C.L., Ma, Q.W., Wang, L., et al. (2009). iPS cells produce viable mice through tetraploid complementation. Nature 461, 86-90.

128. Ohi, Y., Qin, H., Hong, C., Blouin, L., Polo, J.M., Guo, T., Qi, Z., Downey, S.L., Manos, P.D., Rossi, D.J., et al. (2011). Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nat Cell Biol 13, 541-549.

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

130. Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol 28, 848-855.

131. Hussain, T., and Mulherkar, R. (2012). Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. Int J Mol Cell Med 1, 75-87.

132. Caliskan, M., Cusanovich, D.a., Ober, C., and Gilad, Y. (2011). The effects of EBV transformation on gene expression levels and methylation profiles. Human molecular genetics 20, 1643-1652.

133. Research, C.I.f.M. (2016). BIOREPOSITORIES. In. (

134. Group, I.H.W. (2016). Reference Panels. In. (

135. *Austria, B.a.B.R.R.I.* (2016). Catalogue of Austrian Biobanks. In. (

136. (2016). Genetic Repositories Australia. In. (

137. (2016). Taiwan Han Chinese Cell and Genome Bank. In. (

138. Nam, H.Y., Shim, S.M., Han, B.G., and Jeon, J.P. (2011). Human lymphoblastoid cell lines: a goldmine for the biobankomics era. Pharmacogenomics 12, 907-917.

139. Disorders, T.N.C.f.C.G.R.o.M. (2016). The NIMH Stem Cell Center. In. (

140. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol 165, 1328-1335.

141. Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., et al. (2002). Intra- and interspecific variation in primate gene expression patterns. Science 296, 340-343.

142. Khaitovich, P., Tang, K., Franz, H., Kelso, J., Hellmann, I., Enard, W., Lachmann, M., and Pääbo, S. (2006). Positive selection on gene expression in the human brain. Curr Biol 16, R356-358.

143. Fraser, H.B., Khaitovich, P., Plotkin, J.B., Pääbo, S., and Eisen, M.B. (2005). Aging and gene expression in the primate brain. PLoS Biol 3, e274.

144. Waddington, C.H. (1942). Canalization of Development and the Inheritance of Acquired Characters. In., pp 563-565.

145. Debat, V., and David, **P.** (2001). Mapping phenotypes: canalization, plasticity and developmental stability. In. (TRENDS in Ecology & Evolution.

146. Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. Dev Suppl, 135-142.

147. Kalinka, A.T., and Tomancak, P. (2012). The evolution of early animal embryos: conservation or divergence? Trends Ecol Evol 27, 385-393.

148. Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 18, 675-685.

149. Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 20, 1131-1139.

150. Chu, L.F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D.T., Choi, J., Kendziorski, C., Stewart, R., and Thomson, J.A. (2016). Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol 17, 173.

151. Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N.K., Macaulay, I.C., Marioni, J.C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289-293.