THE UNIVERSITY OF CHICAGO

DEVELOPING INDUCED PLURIPOTENT STEM CELL MODELS FOR PRIMATE 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

COMMITTEE ON HUMAN GENETICS

BY
BRYAN PAVLOVIC

CHICAGO, ILLINOIS
JUNE 2018

Table of Contents

List of Figures

List of Tables [1]

1. Note: Due to the large size of some tables, the tables have been provided in a supplementary file accompanying the dissertation. In such cases, the page number provided below directs the reader to a table's caption.

ABSTRACT

Comparative studies in primates are extremely restricted because we only have access to a few types of cell lines from non-human apes and to a limited collection of frozen tissues. In order to gain better insight into regulatory processes that underlie variation in complex phenotypes, we must have access to faithful model systems for a wide range of tissues and cell types. To facilitate this, we have generated a panel of fully characterized chimpanzee (*Pan troglodytes*) and human induced pluripotent stem cell (iPSC) lines derived from fibroblasts of healthy donors. All lines are free of integration from exogenous reprogramming vectors, can be maintained using standard iPSC culture techniques, and have proliferative and differentiation potential similar to human and mouse lines. To begin demonstrating the utility of comparative iPSC panels, we collected RNA-seq data and methylation profiles from the human and chimpanzee iPSCs and their corresponding fibroblast precursors. In order to demonstrate the utility of this panel in studying dynamic developmental processes we collected RNA-seq for a four day time course. We observed conservation of this early developmental period with a reduction of variance shared by both species at the onset of differentiation. In order to determine the fidelity of iPSC derived cell types to a primary tissue, we collected RNA-seq for cardiomyocytes derived from our human and chimpanzee iPSC panel as well as from postmortem heart tissue from both species. We found that while iPSC-derived cardiomyocytes are still in a fetal state, they are able to recapitulate many of the interspecies gene expression differences identified in adult heart tissue, indicating they are a good model to study inter-primate gene regulatory differences. The work contained in this thesis will help realize the potential of iPSCs, and in combination with genomic technologies, transform studies of comparative evolution.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Heritable mutations altering the temporal, spatial or steady state levels of gene expression are likely to underlie the molecular basis for many human-specific traits [58]. Evidence from model systems along with a more thorough understanding of the molecular mechanisms determining transcription have largely supported this hypothesis. For example, it has been hypothesized that human-specific gene expression patterns in the brain might underlie functional, developmental, and perhaps cognitive differences between humans and other apes [165, 92]. To study these regulatory differences between the species, the field has used several approaches, including the use of post mortem frozen tissues, cell lines, and model organisms. The use of primate cell lines and frozen tissues has allowed us to study the genetic architecture of gene regulation. At the same time, because of the limitations of these approaches, it is difficult to study inter-species variation in gene regulation [119]. For example, it is rare to collect a large number of tissue samples from the same donor (particularly in non-human primates). Therefore, it is difficult to investigate population-level patterns of gene regulation in multiple cell types or tissues that are derived from the same genotype (same donor) across species. Furthermore, post-mortem tissues cannot be staged and are often subject to high environmental variances. This variance results in sample heterogeneity that is difficult to measure, and therefore account for, during an experiment. Since frozen tissues are static and non-renewable, they not amenable to experimental perturbation. As a result, we cannot comparatively study the population dynamics of primate gene regulation. A few studies have chosen to sidestep this limitation by using model organisms in an attempt to recapitulate inter-primate differences in gene regulation. In this approach, the phenotypic impact of differences in gene regulation between humans and non-human primates are studied with high

spatial and temporal resolution in a model species, such as the mouse [40, 104, 105, 126]. These studies are useful and often informative, but when model organisms are used to recapitulate gene regulatory differences between primates, the inference about function requires one to make a critically important assumption, which is typically not tested. Namely, one must assume that the effects of gene regulatory changes on complex phenotypes are identical in model organisms and in humans. The caveats associated with using a model organism to study the phenotypic effects of regulatory differences between primates notwithstanding, until recently, it is not clear that there was an alternative approach. In order to gain true insight into regulatory processes that underlie variation in complex phenotypes, we must have access to faithful model systems for a wide range of tissues and cell types. In other words, to utilize comparative functional approaches to comparatively study the genetic architecture of complex phenotypes in humans and other apes, a new approach is needed. Recent technological developments in the generation and differentiation of induced pluripotent stem cells (iPSCs) provide an enticing option to comparatively study primate biology.

1.2 iPSC based systems for comparative primate studies

Human embryonic stem cells (hESCs) are a cell type derived from the inner cell mass of fertilized blastocysts that are able to differentiate into any cell type in the body and thus considered pluripotent. hESCs have long been used as a tool in the study of genetic mechanisms dictating early developmental trajectories. The destruction of the blastocyst in the process of hESC isolation has led to ethical and practice issues surrounding the use of hESCs in scientific research. Due to recent restrictions on the use of great apes in biomedical research, the issues surrounding ESC technology are further confounded and preclude their application to great apes. With the advent of human induced stem cell technology[137, 138] it is now possible to generate cells that are highly similar to hESCs by transient over-expression of pluripotency associated transcription factors. Initially cell lines were generated using retro-

viral integration of *POU5F1*,*SOX*,*KLF4*, and *MYC* (often referred to as Yamanaka factors after the lab that discovered their utility in generating iPSC lines) into dermal fibroblast lines from both mice and humans. As the field matured, more sophisticated methods using non-integrating approaches have been developed[97], resulting in the production of higher-quality integration free cell lines. iPSC lines have been shown to be capable of essentially unlimited proliferation potential with minimal accumulation of mutations over time, thus allowing for the production of a consistent and renewable stream of terminal cell types through staged differentiation protocols. In chapter 2, I establish a novel primate model system employing iPSC technology. Developing such an iPSC-based system for comparative genomic studies in primates allows researchers to compare regulatory pathways and complex phenotypes in humans and our close evolutionary relatives using appropriate models for different tissues and cell types. This will be a powerful resource with which to examine the contribution of changes in gene regulation to human evolution and diversity.

1.3 Fidelity of pluripotent stem cell models to *in vivo* dynamic processes and cell types

Primate iPSC panels are a particularly attractive system for comparative studies of early development. Utilizing timecourse differentiation protocols, we can examine the context dependent nature of gene regulation, as well as the temporal roles of gene expression as different cell types and developmental states are established [73]. However, the biological relevance of differences observed in a comparative iPSC based model hinges on the assumption that iPSCs and differentiation protocols are able to faithfully represent *in vivo* dynamic developmental processes and cellular counterparts. In order to assay comparative iPSC based models in their ability to faithfully recapitulate dynamic *in vivo* developmental processes, I choose to study endoderm development, with the expectation that germ layer establishment will be largely conserved throughout mammals [176]. In chapter 3, I examine the ability of

iPSCs to recapitulate a conserved early developmental gene regulatory trajectory in human and chimpanzee iPSC-derived cell lineages. My rationale is that if I can demonstrate the fidelity of the iPSC model in this context under the expectation of conservation, this would provide support for the notion that the iPSC system can be a useful tool for future comparative studies of more diverged dynamic biological processes. While we largely observed a conservation in an early developmental process across species using our comparative iPSC system. There remains a possibility that differentiated cells derived from primate iPSCs do not perfectly recapitulate gene regulatory differences observed within primary tissues between species. Thus, in chapter 4, I comprehensively benchmark the performance of iPSC derived cardiomyocytes in their ability to recapitulate the inter-species gene expression differences observed in post mortem cardiac tissues.

CHAPTER 2

A PANEL OF INDUCED PLURIPOTENT STEM CELLS FROM CHIMPANZEES: A RESOURCE FOR COMPARATIVE FUNCTIONAL GENOMICS

2.1 Abstract[1]

Comparative genomics studies in primates are restricted due to our limited access to samples. In order to gain better insight into the genetic processes that underlie variation in complex phenotypes in primates, we must have access to faithful model systems for a wide range of cell types. To facilitate this, we generated a panel of 7 fully characterized chimpanzee induced pluripotent stem cell (iPSC) lines derived from healthy donors. To demonstrate the utility of comparative iPSC panels, we collected RNA-sequencing and DNA methylation data from the chimpanzee iPSCs and the corresponding fibroblast lines, as well as from 7 human iPSCs and their source lines, which encompass multiple populations and cell types. We observe much less within-species variation in iPSCs than in somatic cells, indicating the reprogramming process erases many inter-individual differences. The low within-species regulatory variation in iPSCs allowed us to identify many novel inter-species regulatory differences of small magnitude.

2.2 Introduction

Comparative functional genomic studies of humans and other primates have been consistently hindered by a lack of samples [119]. In spite of their clear potential to inform our understanding of both human evolution and disease, practical and ethical concerns surround-

1. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, Kagan CL, Burnett JE, Huang CH, Mitrano A et al. 2015. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife 4: e07103. doi: 07110.07554/eLife.07103

ing working with non-human primates have constrained the field to using a limited set of cell types collected in a non-invasive or minimally invasive manner, primarily lymphoblastoid cell lines (LCLs) and fibroblasts. Comparative studies of any other primate tissue have been limited to using post-mortem (typically frozen) materials, thereby precluding most experimental manipulation and yielding primarily observational insights (see [13, 14, 18], for examples). An alternative has been to use model organisms in an attempt to recapitulate inter-primate regulatory differences. The typical approach involves the introduction of sequences of evolutionary interest into a model system, and then searching for spatial or temporal differences in gene expression that can be ascribed to the introduced sequence [34, 40]. This is a difficult and challenging approach and, perhaps as a result, there are still only a handful of well-described examples of human-specific regulatory adaptations in primates [81, 105] and even fewer cases where the underlying regulatory mechanisms have been resolved [104, 117]. While these studies are useful and often informative, they also entail assumptions of functional conservation between the model system and the species of interest that may not necessarily be true [119]. Induced pluripotent stem cells (iPSCs) can provide a viable means of circumventing these concerns and limitations, at least with respect to the subset of phenotypes that can be studied in *in vitro* systems. Reprogramming somatic cell lines to a stable and self-sustaining pluripotent state [137, 138] has become routine practice for human and murine cell lines, but extension to other animals, especially non-human primates, is not yet widespread despite some exceptions e.g. [9, 76]. Instead, the broadest application of iPSCs to date has been the generation of lines derived from patients suffering from a variety of genetic disorders [32, 51, 72, 83, 153], with the dual aims of providing a deeper understanding of disease phenotypes and developing new therapeutic avenues. These cell lines have been shown to display *in vitro* properties corresponding to relevant patient phenotypes observed *in vivo*, both as iPSCs and when differentiated into other pertinent cell types, supporting their utility in clinical applications; more generally, these properties

also highlight the tantalizing flexibility of iPSCs as a means of exploring developmental and cell lineage determination pathways. Thus, the development of an iPSC-based system for comparative genomic studies in primates will allow us to compare regulatory pathways and complex phenotypes in humans and our close evolutionary relatives using appropriate models for different tissues and cell types. This will be a powerful resource with which to examine the contribution of changes in gene regulation to human evolution and diversity. To demonstrate the validity of this approach, we have generated a panel of 7 chimpanzee iPSC lines that are fully characterized and comparable to human iPSC lines in their growth and differentiation capabilities.

2.3 Results

We generated a panel of iPSC lines from seven chimpanzees through electroporation of episomal plasmids expressing *OCT3/4* (also known as *POU5F1*), *SOX2*, *KLF4*, *L-MYC*, *LIN28*, and an shRNA targeting *TP53* [97], as well as an *in vitro*-transcribed *EBNA1* mRNA transcript [27, 50] that promotes increased exogenous vector retention in the days following electroporation. Our chimpanzee panel is comprised of seven healthy individuals (4 female, 3 male, further details on these individuals are given in Supplementary Table 2.1) ranging from 9 to 17 years old. Fibroblasts from 5 of the 7 individuals were purchased from the Coriell Institute for Medical Research, while the remaining two (C6, C7) were derived from 3 mm skin punch biopsies directly collected from animals at the Yerkes Primate Research Center of Emory University (see methods). All chimpanzee iPSC lines described in this publication are available fully and without restrictions to other investigators upon request to the corresponding authors.

2.3.1 Characterizing the chimpanzee iPSCs

The chimpanzee iPSC lines closely resemble human iPSC lines in morphology (Fig. 2.1A;all images shown in main text are from chimpanzee line C4. Similar images of the other lines are available as Supplementary Figs. 2.8-2.12). All lines could be maintained in culture for at least 60 passages without loss of pluripotency or self-renewal capability using standard iPSC culture conditions, both on mouse embryonic fibroblast (MEF) feeder cells and in feeder-free conditions. The genomes of all our lines appeared to be cytogenetically stable; all exhibited normal karyotypes after reprogramming and more than 15 passages in culture, ruling out the presence of gross chromosomal abnormalities (Fig. 2.1B and Supplementary Fig. 2.8). We confirmed nuclear expression of *OCT3/4*, *SOX2* and *NANOG* in all lines by immunocytochemistry (Fig. 2.1C; Supplementary Fig. 2.9). The pluripotent cells also express the surface antigens Tra-1-81 and SSEA4, while cells collected from the center of differentiating colonies expressed SSEA1 at levels comparable to differentiating colonies of human iPSC lines (Supplementary Fig. 2.10). To confirm that the observed expression of pluripotency-associated genes is of endogenous origin, we performed qPCR with primers designed to specifically amplify the endogenous *OCT3/4*, *SOX2* and *NANOG* and *L-MYC* transcripts (Fig. 2.1D; all PCR primers used in this work are listed in supplementary table 2). Indeed, we found no evidence of exogenous gene expression after 10 passages (Supplementary Fig. 2.11), and no traces of genomic integration or residual episomal plasmid retention after 15 passages (Fig. 2.1E). These observations indicate that self-renewal in our chimpanzee iPSC lines is maintained solely through endogenous gene expression.

To confirm pluripotency and test the differentiation capabilities of our lines, we performed a number of assays. First, we generated embryoid bodies from all 7 chimpanzee iPSC lines and assayed their ability to spontaneously differentiate into the three germ layers by immunocytochemistry. All lines spontaneously gave rise to tissues from the three germ layers (Fig. 2.2A; Supplementary Fig. 2.13). Second, we carried out directed differentiations to

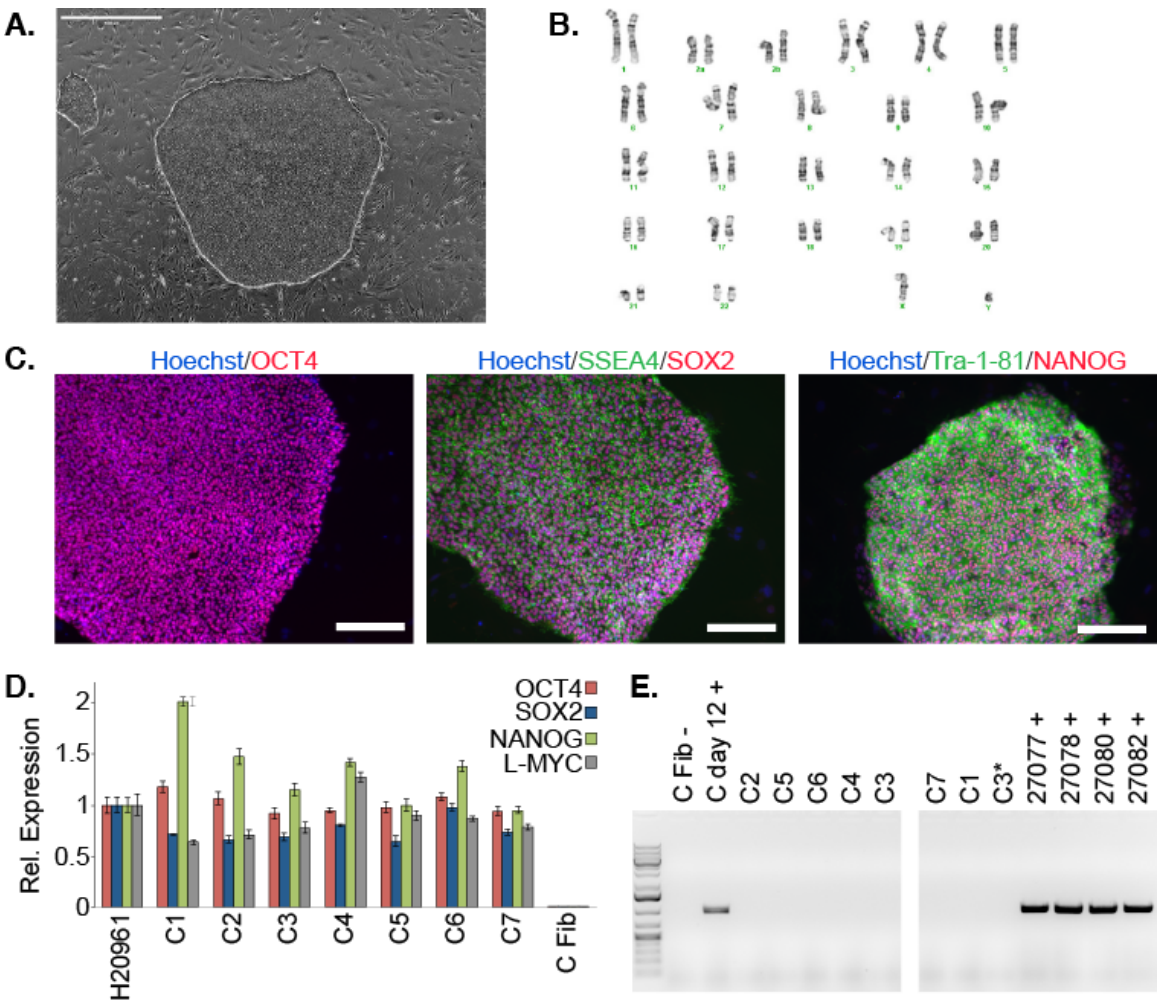


Figure 2.1: **Characterization of chimpanzee iPSC lines.** A. Phase contrast image of representative chimpanzee iPSC line. Scale bar: 1000 μM. B. Representative karyotype from chimpanzee iPSC line after > 15 passages, showing no abnormalities. C. ICC staining of iPSC lines with antibodies for pluripotency markers as indicated. Scale bar: 200 μM. D. Quantitative PCR testing for expression of endogenous pluripotency factors in all 7 chimpanzee iPSC lines. Line H20961 is a male human iPSC line generated in-house used as reference. E. PCR gel showing an absence of exogenous episomal reprogramming factors in all 7 chimpanzee iPSC lines. All PCRs were carried out on templates extracted from passage >15 with the exception of C3651*, which is from passage 2. Fib-is a negative fibroblast control (from individual C8861) prior to transfection, day 12 + is a positive control 12 days after transfection, 27077 + to 27082 + are the plasmids used for reprogramming.

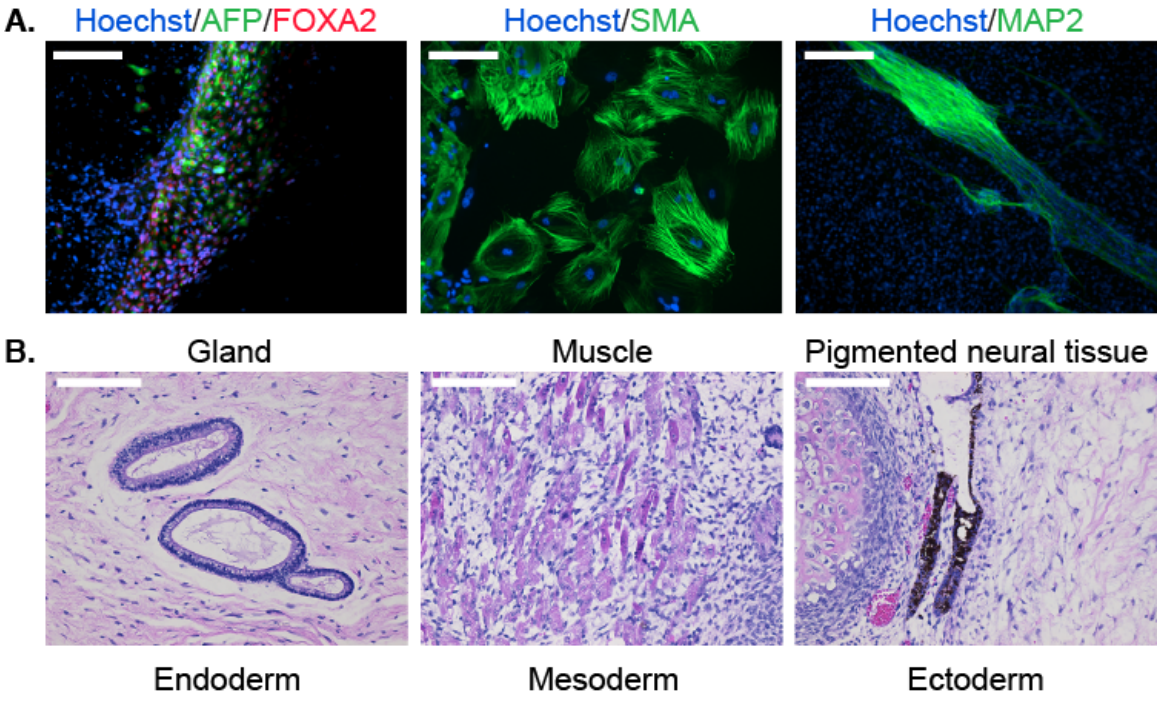


Figure 2.2: **Embryoid body and teratoma assays of iPSC differentiation.** A. ICC staining of differentiated embryoid bodies with antibodies for the three germ layers as indicated. Scale bar: 200μM. B. Histological staining of teratomas derived from iPSC line C4955, showing generation of tissues from all three germ layers. Scale bar: 100μM.

hepatocytes and cardiomyocytes in a subset of the lines using previously published protocols (see Methods, Supplementary Fig. 2.14). Third, we performed teratoma formation assays in four of the lines using Fox Chase SCID-beige and CB17.Cg-$Prkdc^{scid}Lyst^{bg\text{-}J}$/Crl immunodeficient male mice. All four iPSC lines were capable of generating tumours in mice, and all tumours examined contained tissues of endodermal, ectodermal and mesodermal origins (Fig. 2.2B, Supplementary Fig. 2.15). To confirm the chimpanzee origin of these tissues, we extracted and performed Sanger sequencing on mitochondrial DNA from the tumours (Supplementary Fig. 2.16).

Finally we characterized pluripotency in our lines through PluriTest, a bioinformatic classifier that compares the gene expression profiles of new lines to those obtained from a reference set of over 400 well-characterized human pluripotent and terminally differentiated lines [87], modified to accommodate data from both species. All chimpanzee lines

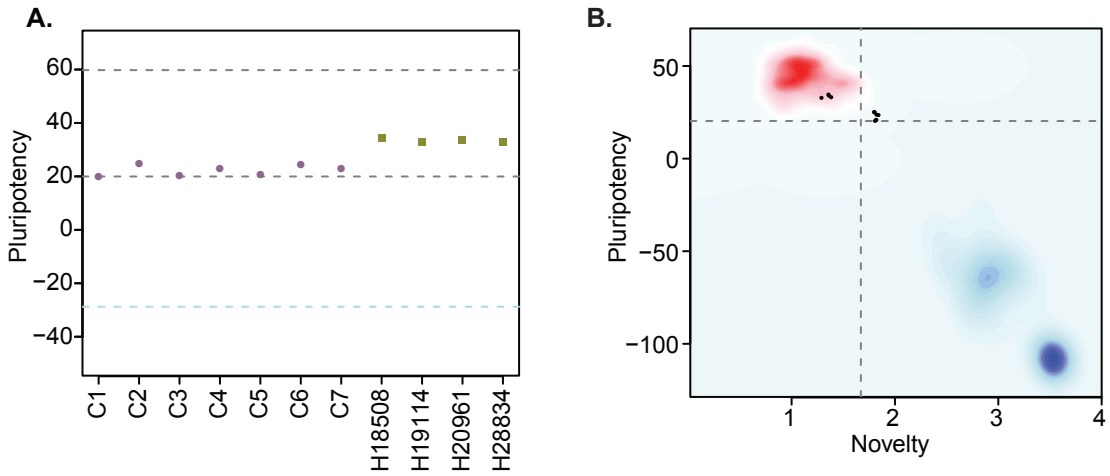


Figure 2.3: **PluriTest**Aa. PluriTest pluripotency scores in the 7 chimpanzee lines and 4 human reference iPSC lines. Purple circles denote chimpanzees; yellow squares, humans. B. PluriTest results after removal of probes not mapping to the chimpanzee genome. All samples in the top left quadrant are human and have satisfactory pluripotency and novelty scores. Samples in the top right quadrant correspond to our chimpanzee iPSC panel, and have consistently high pluripotency yet high novelty scores.

have PluriTest pluripotency scores greater than the pluripotency threshold value of 20 (Fig. 2.3A, Supplementary Table 2.1). We also calculated PluriTest novelty scores for all samples. In human PSCs, novelty values above 1.67 are suggestive of chromosomal duplications or expression of differentiation-associated genes. Human PSCs with high novelty scores are typically either difficult to maintain and expand in culture (because they differentiate spontaneously at a high rate), or cannot be consistently differentiated to all three germ layers. All of our chimpanzee lines had novelty scores above the 1.67 threshold (Fig. 2.3B). However, in contrast to human PSCs with high novelty scores, our chimpanzee lines can be both easily maintained in culture and differentiated into all three germ layer lineages, as demonstrated by the embryoid body and teratoma assays detailed above. We thus hypothesize that the observed high novelty scores are likely driven by inter-species gene regulatory differences that the PluriTest assay, which was trained exclusively on human samples, interpreted as abnormal gene expression.

2.3.2 Interspecies analysis of gene expression and methylation data from iPSCs

To better examine gene expression and regulatory differences between human and chimpanzee iPSCs, we generated genome-wide RNA-sequencing and DNA methylation data (see methods) from all chimpanzee iPSC lines, as well as from 7 human iPSC lines also generated and validated in our laboratory. While all of the chimpanzee iPSCs were derived from fibroblast cell lines (Supplementary Table 2.1), the human iPSCs were derived from both fibroblasts and immortalized LCLs from Caucasian and Yoruba individuals (see Supplementary Table 2.1 for additional details). We designed the comparative study this way in order to demonstrate that regulatory differences between human and chimpanzee iPSCs cannot be explained by technical differences due to culturing conditions or the cell type of the somatic precursor cells used for reprograming. To prevent biases due to genetic divergence between the two species, we chose to restrict our gene expression analyses to a curated set of genes with one-to-one orthology between humans and chimpanzees [13, 15]. Following assessment of quality control metrics (see Methods), we obtained normalized RPKM estimates for 12,171 genes that were expressed in at least 4 iPSC lines from either one of the species (see Methods). We similarly restricted our DNA methylation analyses to a set of 335,307 high quality probes with a high degree of sequence conservation between humans and chimpanzees (as in[47]; see Methods). To examine broad patterns in the data, we used principal component analysis (PCA). We observed clear and robust separation of human and chimpanzee iPSC lines along the first principal component (PC) in both the gene expression and DNA methylation data (Fig. 2.4A,B; regression of PC1 by species; p < 10-13 for the expression data; p < 10-12 for the DNA methylation data). Within the human samples, PC2 appears to be driven by ethnicity, as we observe all Caucasian samples consistently clustering together despite their different cell types of origin ($P = 0.005$ for the association between PC2 and human ethnicity in the expression data, $P = 0.044$ in the DNA methylation data).

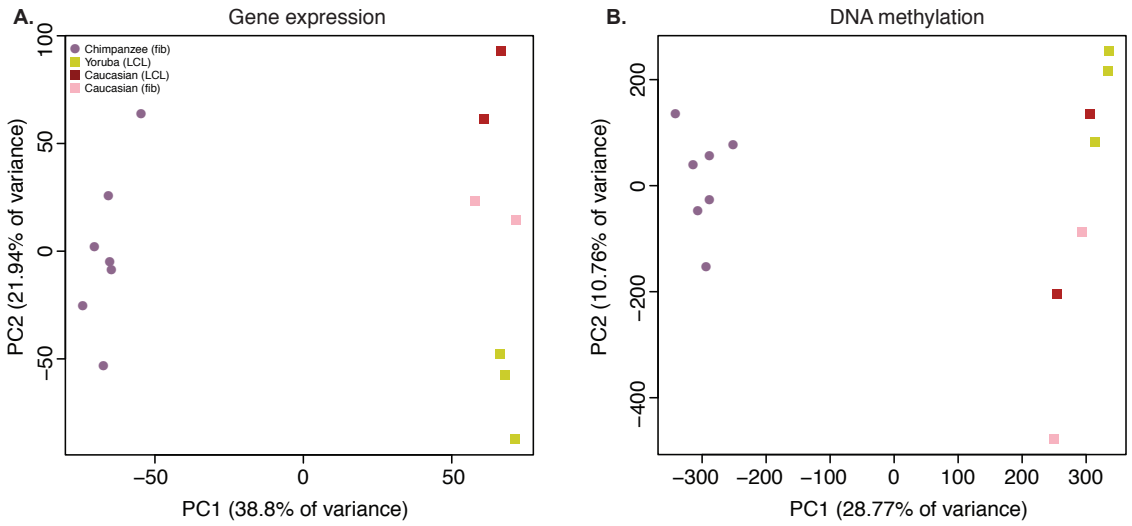


Figure 2.4: **Principal component analysis plots of data from the iPSCs.** A. PCA generated from expression data of 12,171 orthologous genes. B. PCA generated from DNA methylation data measured by 335,307 filtered probes.

We then analysed regulatory differences between the species by first focusing on the gene expression data. At an FDR of 1%, we identified 4,609 genes (37.9%) as differentially expressed (DE) between the iPSCs of the two species (Supplementary Table 2.3; see Methods for details). The majority of DE genes do not exhibit large inter-species fold-change differences in expression levels (Supplementary Figs. 2.18 & 2.19). An analysis of functional annotation of the DE genes reveals that no Gene Ontology Biological Process terms (GO BP; [3]) are significantly overrepresented among these genes at an FDR of 5% (Supplementary Table 2.4), although we identified 123 overrepresented terms if we limit our analysis to the 546 genes with absolute $\log_2$ fold-change difference > 2 (Supplementary Table 2.4). Additionally, we tested for concordance between our list of DE genes and a list of 2,730 genes that were previously classified as DE between human and non-human primate iPSC lines [76]. Given our stringent approach to consider orthologous genes, only 2,081 (76%) genes could be analysed across the two studies. Of these, 1,495 genes are detectably expressed in our lines, and 1,079 (72.2%) are classified as DE between the species in both data sets (a highly significant enrichment; χ^2 $P < 10^{-16}$). Expression trends within these DE genes are

in the same direction in both data sets in 1,060 of cases (98.24%).

Next, we used a similar approach to identify differentially methylated (DM) probes and regions between the iPSCs of both species (see Methods). We identified 63,791 probes that are DM between the two species at an FDR of 1%, 26,554 of which have a mean intergroup βdifference ≥ 0.1, our arbitrary effect size threshold for retaining probes for differentially methylated region (DMR) identification and downstream analyses. Of these, 10,460 probes could be further grouped into 3,529 regions of 2 or more DM probes within 1 kb, which we designated DMRs; (Supplementary Table 2.5); the numbers of probes and regions identified as DM at a range of mean interspecies βthresholds are given in Supplementary Table 2.6. In order to consider the DNA methylation and gene expression data jointly, we focused on a subset of 2,348 DMRs that could be associated with a single Ensembl gene. Overall, these DMRs were associated with 2,141 genes, of which 1,350 were also detectably expressed in the iPSCs, and 558 (41.3%) were classified as DE between the species, a slightly higher proportion than expected by chance alone ($P = 0.1$). We further classified the DMRs as either 'promoter', 'genic' or 'mixed' depending on their position relative to annotated gene transcripts (see Methods) The overall set of DMRs, as well as genic DMRs, are significantly associated with 4 and 79 GO BP terms respectively (FDR $< 5\%$), including terms related to neurogenesis and skeletal system development. Enrichment of several terms related to neurogenesis and skeletal system development is likewise marginally significant amongst promoter and mixed DMRs (Supplementary Table 2.7). However, the subset of inter-species DE genes that are also associated with DMRs are not significantly enriched with annotation for any GO BP or MF terms.

2.3.3 Comparative histone modification data

We used ChIP-seq to characterize the genome-wide distribution of two types of histone modifications (H3K27me3 and H3K27ac) in three of our chimpanzee iPSCs (see Methods).

We compared the chimpanzee data to histone modification data from three human iPSC lines from the Roadmap Epigenomics project (Fig. 2.5). To do so, we downloaded raw sequence files from GEO and processed data from both species using the same pipeline (see Methods). We identified ChIP-seq peaks using MACS or RSEG, as appropriate, and accounted for differences in genome sequence between the species as well as for incomplete power to identify peaks across species (see Methods). To relate the ChIP-seq data to genes (and integrate over data from all peaks that are in proximity to a given gene), we then generated enrichment ChIP scores for a set of previously defined 26,115 orthologous transcription start sites (from: [174]). The enrichment score (see Methods for details, also Supplementary Table 2.8), reflects the ratio of mapped ChIP-seq read counts across all peaks within a 4 kb window centred on an orthologous TSS, relative to the genome-wide read count average after adjusting expectations based on the input control sample. We chose to classify as 'enriched' any region where the mean enrichment score across all three individuals in the species was larger than 1. This cut-off is arbitrary, but we confirmed that our qualitative results are robust by additionally testing enrichment cut-offs of 2, 5, and 10. Using this approach, we first we examined genome-wide patterns of H3K27me3 enrichment in chimpanzee and human iPSCs. Overlap across the two species is considerably higher than expected by chance (Fig. 2.5A, χ^2 $P < 10^{-16}$), but it is somewhat unclear how to interpret this observation with respect to the expectation that human and chimpanzee iPSCs would have similar pluripotency potential. We thus focused on a set of 3,913 genes [66] previously annotated as bivalently modified in human PSCs PSCs-that is, genes known to be associated with both high H3K4me3 and H3K27me3, indicative of a 'poised' or 'primed' state [11]. We expect the vast majority of these genes to also be associated with similar modifications in chimpanzee iPSCs. Only 2,910 of the known bivalent genes were associated with clear orthologous TSSs and could be tested using our comparative H3K27me3 ChIP-seq data. Of these, 306 were not associated with the modification in either species, whereas of the 2,604 genes that were associated with

H3K27me3 in at least one species, 2,368 (90.1%) were enriched for H3K27me3 in both species (Fig. 2.5B, χ^2 $P < 10^{-16}$).

We then examined H3K27ac enrichment patterns in both species. This mark is indicative of active promoters and gene transcription. Overall, we find good agreement between human and chimpanzee genes enriched for H3K27ac, with 95.8% human genes associated with the mark also enriched in chimpanzees (Fig. 2.5C). However, there is a clear excess of genome-wide H3K27ac signal in chimpanzee iPSCs relative to humans, possibly due to an overall more sensitive ChIP enrichment in the chimpanzee samples (Supplementary Figs. 2.20 & 2.21). We proceeded by focusing on a list of 22 core pluripotency transcription factors (drawn from [91, 98]), where we expect to find H3K27ac signal shared across the two species at a higher rate than in the genome-wide data, given the role of these factors in maintaining pluripotency. Due to our stringent requirements for establishing orthology, we were initially able to examine data from 14 of those genes; 11 of which were associated with H3K27ac in both species (Fig. 2.5D)-one of the discrepancies is *REX1* (also known as *ZFP42*), which we discuss further below. We then extended our analysis to include the full set of 22 pluripotency transcription factors regardless of orthology, by testing solely for absence or presence of signal peaks identified by MACS (i.e., without considering enrichment scores; see Methods). We again found a high overlap in H3K27ac enrichment across species, with 15 of the 22 genes associated with H3K27ac enrichment in both species (including the three master regulators of pluripotency, *OCT4*, *SOX2*, and *NANOG*; Supplementary Fig. 2.22). Of the remaining 7 genes, one (*DAX1*) was not found to be associated with H3K27ac in either species, four genes (*ESSRB*, *KLF2*, *KLF4*, and *KLF5*) were associated with H3K27ac only in chimpanzee (although this observation may reflect incomplete power to detect peaks in the human data), and only two genes (*ZFX* and *REX1*) were associated with H3K27ac in human but not in chimpanzee iPSCs.

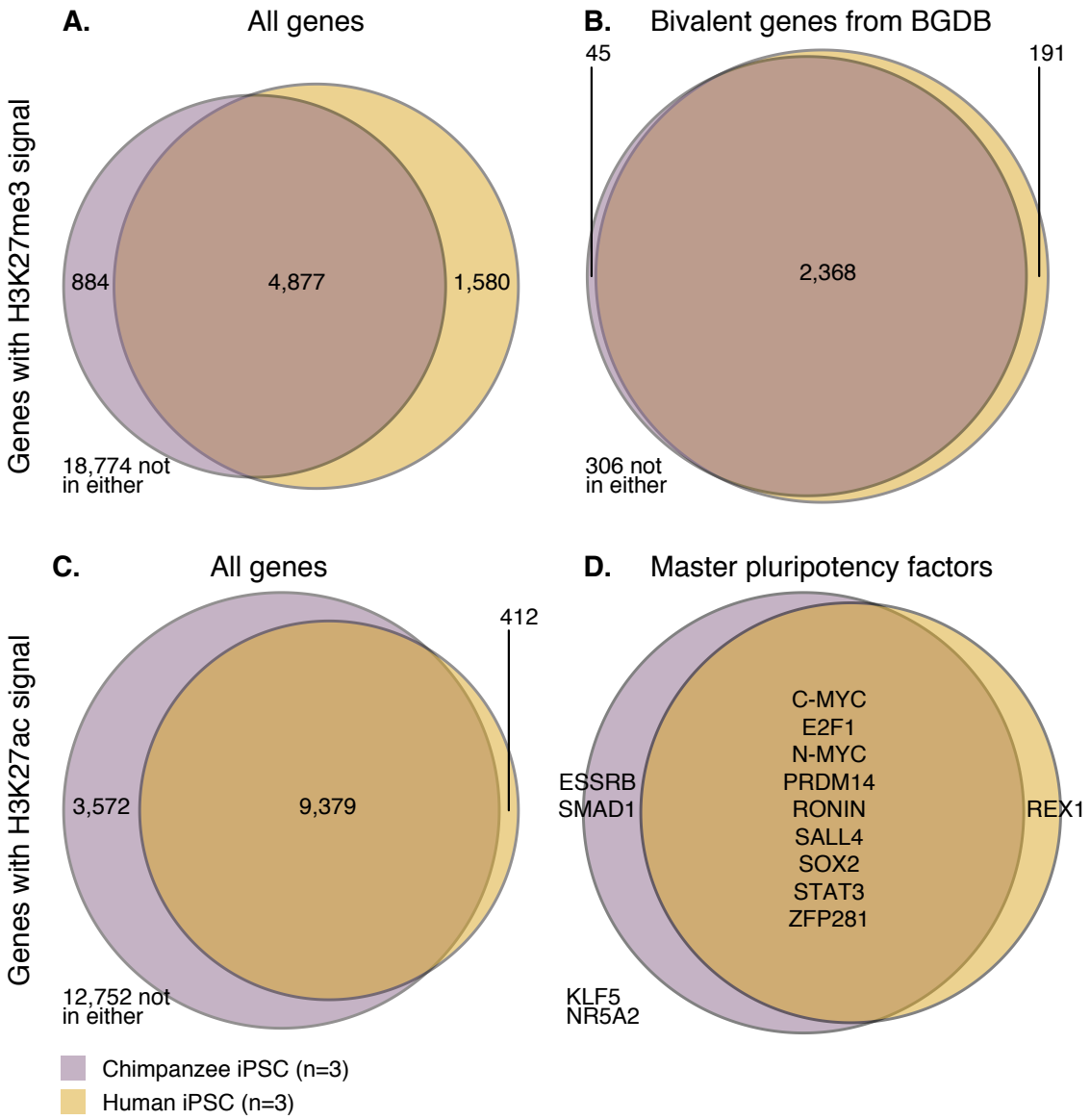


Figure 2.5: **Overlap of H3K27me3 and H3K27ac signal between chimpanzee and human iPSCs at orthologous TSSs.** A. H3K27me3 enrichment in all genes with an orthologous TSS. B. H3K27me3 enrichment in 2,910 genes previously identified as bivalent in human PSCs. C. H3K27ac enrichment in all genes with an orthologous TSS. d. H3K27ac across 14 known pluripotency master regulators with orthologous TSSs.

2.3.4 REX1 may be dispensable for chimpanzee pluripotency

In order to further consider inter-species differences in the core pluripotency regulatory network, we examined expression levels in our chimpanzee and human iPSCs in the same list of 22 core pluripotency TFs described above. Expression values in all iPSC lines are shown in Fig. 2.6A (see also Supplementary Fig. 2.23). Given the stringency of our interspecies analysis approach with respect to unique read mapping, we are unable to calculate RNA-seq-based expression estimates for six of these TFs, including *OCT4* or *NANOG*, both of which have multiple pseudogenes that can confound mapping algorithms (however, as shown in Figure 1D, our qPCR results demonstrate that expression of those 2 genes is similar amongst all chimpanzee iPSC lines, and marginally higher than in our human iPSC control line). Of the 16 TFs with expression data for iPSCs from both species, 4 (*E2F1*, *ESRRB*, *SALL4* and *REX1*) are DE between human and chimpanzee iPSCs at an FDR of 1%. Of these, *ESRRB* and *REX1* are associated with absolute inter-species expression $\log_2$ fold-changes > 1. However, because *ESRRB* is expressed at very low levels across all samples (mean RPKM across all 14 samples = 0.47), we focused our subsequent analyses on *REX1*, which is expressed at low or undetectable levels in 6 of our 7 chimpanzee iPSCs (mean RPKM = 0.667), but at high levels in all human iPSC lines (mean RPKM = 180.58) and a single chimpanzee iPSC, C6 (Fig. 2.6A). Our DNA methylation data is consistent with this gene expression pattern: all 10 probes located in the 5' UTR or up to 1,500 bp upstream from the *REX1* TSS are highly methylated in the six chimpanzee lines (mean βacross all promoter probes = 0.87), but exhibit intermediate or low levels of DNA methylation in all of the human iPSC lines and the *REX1*-expressing C6 line (Fig. 2.6B); the entire region is a DMR (Supplementary Table 2.5). Consistent with these findings, *REX1* is also differentially enriched for H3K27ac signal in the two species - we identified no H3K27ac peaks at the *REX1* TSS in the three chimpanzee lines, which did not include C6 (Fig. 2.5D, Supplementary Fig. 2.22).

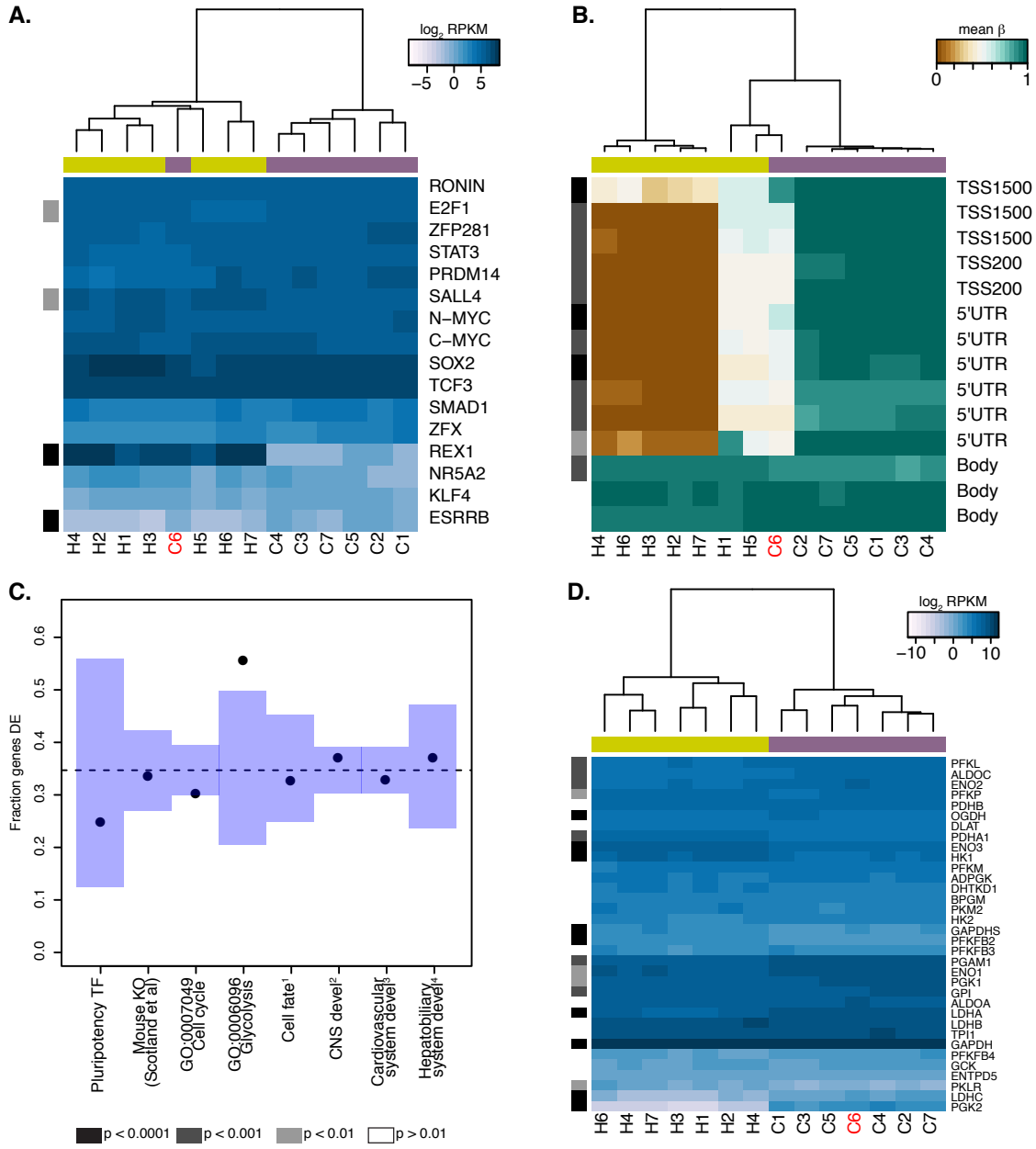


Figure 2.6: ***REX1* may be dispensable for pluripotency in chimpanzee iPSCs.** A. Expression values of 16 core pluripotency transcription factors in all human and chimpanzee iPSC lines. B. Methylation status of 13 CpG sites associated with *REX1* in all human and chimpanzee iPSCs. Location of the probe relative to the gene sequence is indicated along the right hand side. In both panels *REX1*-expressing chimpanzee iPSC line is colored red, significant interspecies differences are indicated along the left-hand side, and purple boxes indicate chimpanzee lines, yellow boxes indicate human lines.

The *REX1* genes codes for a transcription factor present in all placental mammal species, which has long been established as a marker of pluripotency in human and mouse PSCs [20]. Multiple publications have suggested that this gene plays an important role in maintaining pluripotency and inhibiting differentiation into the three primary tissue germ layers [80, 123, 132], with multiple mechanisms of action having been proposed. However, *REX1*-knockout mouse ESC lines can give rise to chimeric animals, and homozygous F2 *REX1* null mice are viable [80], suggesting that *REX1* may not be indispensable for murine pluripotency. In humans, loss of *REX1* expression in ESCs following shRNA knockdown has been associated with a rapid loss of pluripotency, as well as a decrease in glycolytic activity and a lack of observable mature mesodermal structures in teratoma formation assays [132]. To determine the consequences of a lack of *REX1* expression in chimpanzee iPSCs, we considered gene expression data from all human iPSC lines and the 6 chimpanzee iPSC lines that do not express *REX1*. We asked whether there is an excess of DE genes among those thought to be directly regulated by, or downstream of, *REX1* (Fig. 2.6C,D; see Methods), but failed to find enrichment in all categories except for genes associated with GO term BP:0006096, glycolysis, where 19 of 34 testable genes were DE at an FDR of 1% between the two species (/emphP < 0.01 from 100,000 permutations). The direction of this effect ran contrary to previous reports, however, with genes highlighted by [132] as downregulated following *REX1* knockdown, such as *PGAM1* or *LDHA*, having significantly higher expression in chimpanzee iPSCs than in human iPSCS (Fig. 2.6D). Furthermore, the *REX1*-expressing line C6 is not an outlier amongst the other chimpanzee iPSC lines (Fig. 2.6D), suggesting that the observed inter-species regulatory differences cannot be attributed to differences in *REX1* expression between the species. We note that both the teratomas and EBs generated from chimpanzee iPSC lines that do not express *REX1* gave rise to mature structures from all three germ layers similar to those observed in *REX1*-expressing line C6 (Supplementary Figs. 2.13 & 2.13). Furthermore, and consistent with our observations, *REX1* is either absent or expressed

at low levels in one replicate of either of the two retrovirally reprogrammed bonobo (*Pan paniscus*, sister species to chimpanzees) iPSC lines generated by [76] although it is expressed in both replicates of both chimpanzee iPCS lines from the same group (Supplementary Fig. 2.24). Together, these findings suggest that that the variable loss of *REX1* expression in chimpanzee and bonobo iPSCs does not impair pluripotency, and that its regulatory functions of in humans may be being fulfilled in chimpanzee iPSCs by other regulatory mechanisms.

2.3.5 Comparison of iPSCs and other tissues

We collected RNA-sequencing data from all cell lines used to generate both the chimpanzee and human iPSCs (Supplementary Table 2.6). Following quality control and normalisation steps, we obtained RPKM values for 13,147 genes across all 28 iPSC and precursor samples (see Methods). We also obtained DNA methylation profiles from all samples at the same 335,307 probes described above. PCA of both data sets show that the first PC was significantly associated with tissue type in both data sets ($P < 10^{-27}$ for the expression data; $P < 10^{-17}$ for the DNA methylation data; see Fig. 2.7 and Supplementary Table 2.9), while human and chimpanzee samples are separated by species along PC2 ($P = 0.001$ for the expression data; $P < 10^{-4}$ for the methylation data). However, given the absence of chimpanzee LCLs in our dataset, it is not possible to determine whether the separation is driven by tissue type, species, or both. Overall, chimpanzee iPSCs have significantly higher levels of DNA methylation compared to the somatic lines they were generated from ($P < 10^{-15}$; Supplementary Fig. 2.27), an observation that extends to all genomic features we tested (Supplementary Fig. 2.28); similar observations have been previously made in human PSCs [17, 90]. Remarkably, both DNA methylation and gene expression levels in iPSCs are relatively homogeneous within species, far more so than in their corresponding precursor cells (Fig. 2.6B,D; $P < 10^{-14}$ when comparing overall pairwise distances within all chim-

panzee iPSCs and within all chimpanzee fibroblasts in the methylation data; $P < 10^{-9}$ for the same comparison in the gene expression data). DNA methylation levels in iPSCs also have significantly reduced coefficients of variation relative to their precursor lines (range of CVs for chimpanzee iPSCs = 0.78 - 0.80, for chimpanzee fibroblasts = 0.87 - 0.90; $P < 10^{-06}$). We observed the same pattern in the human data, although in this case the multiple somatic origins of the cell lines of origin contribute to the higher level of variation.

We then performed analyses of gene expression and DNA methylation differences in the combined iPSC and somatic precursor dataset. First, we carried out a comparison of the iPSCs and the precursor cells within each species (see Methods) and classified 9,235 genes as DE between chimpanzee fibroblasts and the corresponding iPSCs. In humans the number of DE genes is 7,765 if we consider all iPSC lines and their somatic precursors, 8,087 if we only consider those derived from LCLs (n = 5), and 5,489 if we only consider those derived from fibroblasts (n = 2; Supplementary Table 2.10). Similarly, we identified 18,029 DMRs between chimpanzee fibroblasts and iPSCs, and 12,078 DMRs between all human somatic precursors and all human iPSCs (Supplementary Table 2.11 - 2.12). No GO categories are significantly overrepresented in any of these data sets. Next, we focused on a comparison of inter-species differences in gene expression and DNA methylation levels across cell types. Following joint normalisation and modelling of data from all samples (see Methods), we classified 5,663 genes as DE between the chimpanzee precursor fibroblasts and the collection of human precursor LCLs and fibroblasts, as well as 84,747 DM probes and 9,107 DMRs (always at an FDR of 1%). Most of these regulatory differences, however, reflect variation across cell types rather than across species (6,324 genes and 70,312 probes are DE or DM between the human fibroblasts and LCLs, respectively). We thus considered only data from the fibroblast precursors in the two species. Only 2 of the human iPSCs were reprogrammed from fibroblasts, leading to a loss in power; we were nonetheless able to identify 1,236 DE genes and 25,456 DM probes between human and chimpanzee fibroblasts, and 1,118 DE genes

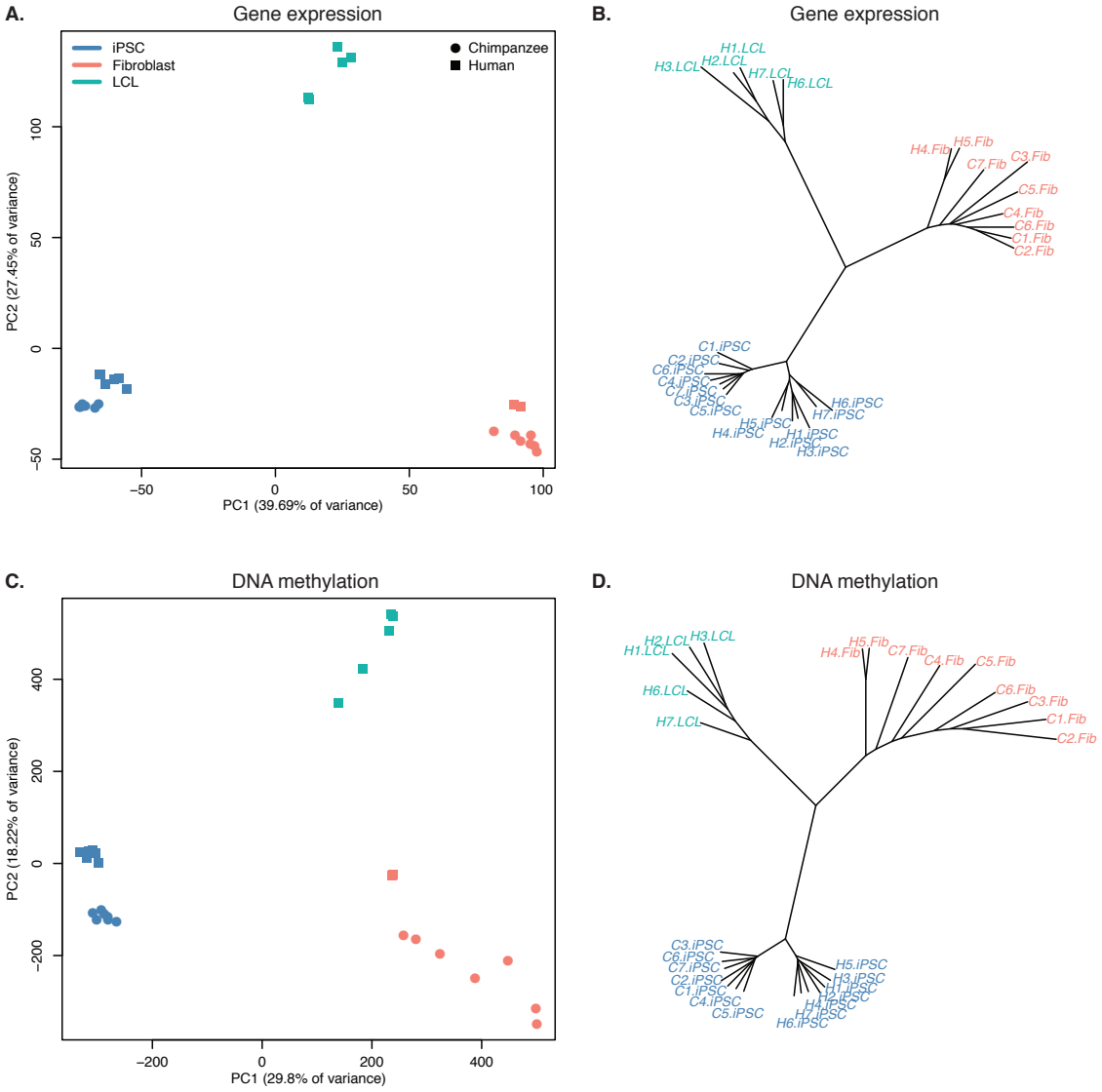


Figure 2.7: **Relationships of iPSCs to their precursors.** A. PCA of gene expression data from all iPSCs and their precursor cell lines. B. Neighbour-joining tree of Euclidean distances between all samples generated based on the gene expression data. C. PCA of DNA methylation data from all iPSCs and their precursor cell lines. D. Neighbour-joining tree of Euclidean distances between all samples generated based on the DNA methylation data.

and 16,392 DM probes between the corresponding iPSCs of the two species. None of these gene sets were significantly enriched for functional annotations using GO BP terms. Although the overlap of inter-species DE genes and DM probes between the iPSCs and the precursors is considerable (13.6% of DE genes and 11.8% of DM probes), a large number of regulatory differences are only observed between the iPSC lines of the two species (Supplementary Fig. 2.29). This observation is robust with respect to different approaches to normalising and modelling the data (Supplementary Fig. 2.30), strongly suggesting that many of the differences we observe between our chimpanzee and human iPSC lines may be intrinsic features of the pluripotent state in these two species.

2.3.6 Discussion

iPSCs have the potential to transform our understanding of the biology of non-model organisms and facilitate functional comparative studies. To this end, we have generated a panel of 7 fully characterized chimpanzee iPSCs. All lines are capable of spontaneously giving rise to the three tissue germ layers *in vitro* and *in vivo* and meet all currently established criteria for pluripotency. The chimpanzee iPSC lines provide a tantalising avenue for investigating how changes in gene expression and regulation underlie the architecture of complex phenotypic traits in humans and our closest living relatives [77, 119]. In particular, we believe that through the use of directed differentiation protocols, functional studies could be performed in cell types where strong *a priori* hypotheses support a role for selective pressure underlying inter-species divergence (e.g. liver, heart, kidney [13, 14]). In that sense, we hope that this panel of cell lines will be a useful tool to researchers interested in overcoming current limitations of comparative studies in primates. To that purpose, all chimpanzee iPSC described in this publication the panel are available fully and without restrictions to other investigators upon request to the corresponding authors. Other groups have previously generated pluripotent stem cells from primates [9, 26, 36, 49, 71, 96, 142, 160, 161, 162]. Indeed, a recent

publication [76] reported the generation of two chimpanzee and two bonobo (*Pan paniscus*) iPSC lines through the use of retroviral vectors. However, in the course of our work we have found that retroviral vector silencing in chimpanzee iPSCs was not as stable as in human iPSC lines generated at the same time using the same method (see Methods and Supplementary Fig. 2.12). Our use of episomal vectors circumvents this problem, and more broadly the problems of both random exogenous gene reactivation and disruption of the host genome through retroviral integration [131]. More generally, while the sum total of primate PSC generation efforts so far has resulted in a sizable number of lines being established from various donors and species these have been generated through various reprogramming protocols and source cell types. We have generated iPSCs from a panel of seven individuals using a consistent protocol and cell type of origin. Given the panel size, it is powerful enough to robustly detect inter-species differences in gene expression, splicing and regulation. The fact that our panel contains both female and male lines also allows for future studies of sex-specific differences in gene expression in various cell types. Indeed, we have previously shown that this can be accomplished using as few as six individuals from each species [13]. Beyond its future applications, however, our panel has already yielded insights into the pluripotent state in chimpanzees and humans. On the one hand, both at the transcriptional and epigenetic level, our iPSCs are remarkably homogeneous both within and between species, significantly more so than their precursors cells. This finding aligns with our current understanding of the reprogrammed pluripotent state as a complex, highly regulated state [54], deviations from which are likely to result in loss of pluripotency and lineage commitment. Additional support for this notion was provided by the strong overlap in H3K27me3 signal between the two species, especially in known bivalent genes. It is remarkable that we have been able to observe this considerable conserved chromatin signature despite the obvious confounding technical batch effect in these comparative data. On the other hand, we were also able to identify over 4500 genes that are DE between human and chimpanzee iPSCs, as well as

over 3500 DMRs between the two species. These numbers are greater than what has been previously observed in comparisons of other tissues across humans and chimpanzees with similar sample sizes ([14, 13]). We believe that the reasons for this difference are likely to primarily stem from increased power to detect DE genes and DMRs in our sample relative to previous work. Given the small amount of intra-species variability we observed in RNA-seq and methylation relative to other tissues, we expect to have greater power to detect small, statistically significantly inter-species differences that would have been missed by studies that consider more variable tissue samples. This notion is supported by the fact that the fraction of genes with log FC < 2 we detect as DE between human and chimpanzee iPSCs is greater than in other comparison we have performed with any other tissue (Supplementary Fig. 2.19). Though small in magnitude, we expect that a subset of these regulatory differences may be biologically relevant (e.g., we find that inter-species regulatory differences in methylation levels are enriched in regions associated with developmental processes; 2.4-2.5). We specifically highlighted an inter-species difference in *REX1* expression levels. This gene is considered an indispensible pluripotency marker in human and mouse PSCs, but our observations suggest that it may not be the case in chimpanzees. Although only one chimpanzee iPSC line expresses *REX1*, we were unable to identify any systematic differences between our human and chimpanzee iPSCs that would indicate a reduction in pluripotency. We also examined *REX1* expression levels in 73 human iPSC lines generated in-house from Caucasian individuals using the Illumina HT12v4 array (Supplementary Fig. 2.25). All lines had PluriTest pluripotency scores > 20, yet 3 of 73 lines (4.1%) showed levels of *REX1* expression that were indistinguishable from background signal, suggesting that *REX1* may not be expressed in these individuals despite their high pluripotency scores. We also examined methylation status at the *REX1* locus in previously published human ESCs and iPSCs from [175], and found that although all ESC lines examined exhibited consistent levels of low methylation at the *REX1* promoter, human iPSC lines analysed in exhibited either hemi-

or hyper-methylated *REX1* promoter regions (Supplementary Fig. 2.26). In the absence of publicly available *REX1* expression data from either of the hiPSC lines with hypermethylated promoters we cannot be certain that the gene is not expressed in these lines, but the combination of these findings with our observations above and previous literature suggest that *REX1* may be important in regulation and maintenance of pluripotency in ESCs, but not necessarily so in iPSCS. Additionally, in chimpanzees, the *REX1* gene has undergone multiple deletions and insertions relative to the human sequence, most significantly a 647 bp insertion in its first intron, and two insertions in the 3' UTR region of approximately 300 bp each that may disrupt the local regulatory landscape; the gene has also been duplicated, with a second copy retrotransposed into chromosome 14; none of these changes are shared with gorillas or orangutans. Although it is currently unclear whether some or all of these changes are also present in the bonobo, these findings might explain why we observed low or no *REX1* expression more frequently in chimpanzees than in humans, and suggests that the gene may not be necessary for maintaining pluripotency in the *Pan* lineage. PSCs have been used to study developmental pathways *in vitro* (for example, [100, 108, 152, 164]). Although optimization of existing differentiation protocols will likely be necessary for application in the chimpanzee system, our panel of iPSCs makes it possible to carry out comparative developmental studies between humans and chimpanzees, and firmly test the hypothesis that changes in gene regulation and expression, especially during development, underlie phenotypic differences between closely related species, especially primates [19, 24, 25, 53, 58]. In addition, we should be able to recreate and test the effect of inter-species regulatory changes in the correct cell type and species environment, enabling studies that cannot otherwise be performed in humans and non-human primates. The use of panels of iPSCs including lines from both humans and non-human primates will thus allow us to gain unique insights into the genetic and regulatory basis for human-specific adaptations.

2.4 Methods

2.4.1 Isolation and culture of fibroblasts

All biopsies and animal care were conducted by the Yerkes Primate Research Center of Emory University under protocol 006-12, in full accordance with IACUC protocols. Skin punch biopsies (3 mm) were rinsed in DPBS containing Primocin (Invivogen) and penicillin/streptomycin (Pen/Strep, Corning) and manually dissected into 10-15 smaller pieces. The tissue was digested in 0.5 % collagenase B (Roche) for 1-2 hours until cells were released from the extracellular matrix. Dissociated cells were pelleted by centrifugation at 250 x g, and the supernatant was spun a second time at 700 x g to pellet any cells that had not been completely released from the extracellular matrix. Cell pellets were resuspended in a 1:1 mixture of -MEM and F12 (both from Life Technologies) supplemented with 10% FBS (JR Scientific), NEAA, GlutaMAX (both from Life Technologies), 1% Pen/Strep, 64 mg/L L-ascorbic acid 2-phosphate sesquimagnesium salt hydrate (Santa Cruz Biotech) and Primocin. Cells were plated in a single well of a 6-well plate coated with 4 $\mu g/cm^2$ human fibronectin (BD Biosciences) and 2 $\mu g/cm^2$ mouse laminin (Stemgent). Cultures were grown at 5% CO_2/5% O_2 until confluent and then split using 0.05% trypsin. For routine passaging cells were cultured at 5% CO_2 and atmospheric oxygen in primate fibroblast media, which is the same as plating media but does not contain F12 base media.

2.4.2 Generation of retrovirally-reprogrammed iPSC lines (a failed attempt)

We initially attempted to generate lines by retroviral transduction using through transfection with pMXs- vectors encoding the human *OCT3/4*, *SOX2*, *KLF4*, *L-MYC* and *NANOG* sequences (Addgene plasmids 17217, 17218, 17219, 26022 and 18115) as well as vectors encoding the MSCV-VSV.G envelope protein (Addgene plasmid 14888) and MSCV gag-pol

(Addgene plasmid 14887). 15 μg of each vector was transfected into 293FT cells (Life Technologies) using Lipofectamine 2000 (Life Technologies) as directed by the manufacturer. We collected virus-containing supernatant from the 293FT cells 48 and 72 h after transfection and immediately used this viral media to transduce chimpanzee fibroblasts, alongside 10 μg/mL of polybrene (Sigma Aldrich H9268). To aid viral penetration, we centrifuged the cells at 1800 RPM for 45 minutes following each transduction. 24 h after the second transduction, we replaced the viral media with A-MEM + 10% FBS, NEAA and Glutamax. Transduced fibroblasts were allowed to recover for a further 2 days and then seeded on γ-irradiated, CF-1-derived mouse embryonic fibroblasts (MEF) at a density of 10,000 cells/cm^2, and maintained in hESC media (DMEM/F12 supplemented with 20% KOSR, 0.1 mM NEAA, 2 mM GlutaMAX, 1% Pen/Strep, 0.1 mM BME and 25 ng/mL human bFGF) supplemented with 0.5 mM valproic acid (Stemgent) until day 14. We obtained iPSCs from 5 chimpanzees by using this protocol. Yet, when we performed quality control and pluripotency checks on these lines we found that the exogenous transfected genes were still expressed (Supplementary Fig. 2.12). Pluripotency in these lines could not be maintained exclusively through endogenous expression. We discarded all 5 lines and proceeded with a different reprograming strategy as detailed below.

2.4.3 Generation of episomally-reprogrammed iPSC lines

Fibroblasts were grown at 5% CO_2/atmospheric O_2 in primate fibroblast media until 70-80% confluence and released by trypsinisation for transfection. 1.510^6 cells were transfected with 1.5 μg per episomal vector containing the following genes: *OCT3/4*, *SHp53*, *SOX2*, *KLF4*, *LIN28*, and *L-MYC* (Addgene plasmids 27077, 27078, 27080 and 27082; [97]). To boost the initial retention of vectors following transfection, 3 g of *in vitro* transcribed ARCA capped/polyadenylated *EBNA1* mRNA was cotransfected with the vectors (see below). Transfected cells were seeded at 15,000/cm^2 on tissue culture plates precoated with 1

g/cm^2 vitronectin (Stemcell Technologies). Cells were grown in Essential 8 media (made in house as previously described in [27]) without *TGFβ1*, supplemented with 0.5 mM sodium butyrate (NaB, Stemgent)) and 100 nM hydrocortisone (Sigma Aldritch). Hydrocortisone was used between days 1-12, or until cell density exceeded >70% confluence. At day 12, cells were detached using TrypLE (Life Technologies) and replated at a density of 5,000 cells/cm^2 on cell culture dishes precoated with 0.01 mg/cm^2 (1:100) of hESC-grade Matrigel (BD Sciences) and grow in Essential 8 media without *TGFβ1* or NaB. Colonies began to form at days 18-22 and were picked between days 24-30 onto dishes coated with γ-irradiated CF-1 derived MEF and subsequently grown in hESC media (as described above) supplemented with 100 ng/mL human bFGF (Milteny Biotech). Clones were routinely split using Rho-associated kinase (ROCK) inhibitor Y27632 (Tocris) at a concentration of 10μM. Cells were migrated to 1:100 hESC Matrigel (BD Sciences) and maintained on Essential 8 media after a minimum of 15 passages on MEF. Feeder free cells were passaged using EDTA-based cell release solution as in [27].

2.4.4 Generation of EBNA1 mRNA

To generate a template for *in vitro* transcription, an *EBNA1* template was designed using the wild type HHV4 *EBNA1* as a reference sequence (NCBI accession YP_ 401677.1). The reference sequence was modified by replacing the GA repeat region and domain B (amino acids 90-375) with a second, tandem, chromatin-binding domain (domain A, amino acids 27-89), similar to what was done by [50]. The nuclear localization signal (amino acids 379-386) was removed and replaced with the sequence GRSS. Using the amino acid sequence as the starting template, the corresponding DNA sequence was generated by reverse translation and optimized for expression in human cell lines using Genscript's OptimumGene codon algorithm. This sequence was synthesized by Genscript and provided in the pUC57 cloning vector; the *EBNA1* coding sequence was subcloned into pcDNA3.1+ (Life Technologies)

using the restriction enzymes BamHI and HindIII. Capped and poly(A) mRNA transcripts were generated using the mMESSAGE mMACHINE T7 ULTRA kit (Life Technologies) with 1 g of BamHI linearized pcDNA3.1+*EBNA1* as the template. The plasmids encoding the wild type and modified *EBNA1* sequences have been deposited to Addgene as plasmid ID# s 59199 and 59198 for the wild type and modified sequences respectively.

2.4.5 iPSC characterization

iPSC colonies were cultured on MEF for 4-6 days and fixed using PBS containing 4% PFA (Santa Cruz BioTech) for 15 minutes at room temperature. After rinsing with PBS, fixed cells were blocked and permeabilized for one hour in PBS containing 0.3% triton and 5% BSA. Primary antibodies: OCT3/4 (SC-5279), SOX2 (SC-17320), NANOG (SC-33759), SSEA-4 (SC-21704), and Tra-1-81 (SC-21706), all from Santa Cruz BioTech, were diluted 1:100 in blocking solution. Fixed cells were incubated with the primary antibody solution overnight on a rocker at 4°C. After washing out the primary antibody solution, fixed cells were incubated with secondary antibodies (labeled with either Alexa-488 or Alexa-594, 1:400, Life Technologies) diluted in blocking for 1 hour on a rocker at room temperature. Nuclei were counterstained using 1 μg/mL Hoechst 33342 (ThermoFisher). All fluorescence imaging was conducted using an AMG EVOS FL (Life Technologies).

2.4.6 Quantitative PCR for endogenous and exogenous gene expression

RNA was extracted using Qiagen RNA miniprep columns from cell pellets collected from fibroblasts, day 7 post transfection and feeder free (Matrigel and Essential 8) iPSC lines at passage 10 or higher for both the retroviral and episomal reprogrammings; 1 μg of total RNA was reverse transcribed using the Maxima first strand cDNA synthesis kit (Thermo Scientific). Quantitative PCR was performed using a 1:96 dilution of cDNA and SYBR Select master mix (Life Technologies) with both forward and reverse primers at a concentration of

0.2μM.). Data was collected and analysed using the Viia7 (Life Technologies). Primer sequences are shown in 2.2, exogenous gene expression melt curves are shown in Supplementary Fig. 2.12.

2.4.7 Generation of embryoid bodies and immunofluorescence

Colonies growing on MEF were detached using Dispase/Collagenase IV (1 mg/ml each; both from Life Technologies) in DMEM/F12 and grown as a suspension culture on low adherent plates using hESC media without bFGF. After one week of suspension growth, cells were transferred to 12 or 24-well plates coated with 0.1% gelatin and grown in DMEM supplemented with 20% FBS, 0.1 mM nonessential amino acids, 2 mM GlutaMAX, 1% Pen/Strep and 64 g/mL L-Ascorbic acid 2-phosphate sesquimagnesium salt hydrate. Embryoid bodies were grown for 1-2 weeks prior to fixation and immunofluorescence staining. Cultures were fixed and stained as described above using the following antibodies: AFP (1:200, SC-130302, Santa Cruz Biotech), FOXA2 (1:200, SC-6554, Santa Cruz Biotech), -smooth muscle actin (1:1500, CBL171, Millipore) and MAP2 (1:200, sc-20172 and sc-74420, Santa Cruz Biotech).

2.4.8 Integration analysis

To test for genomic integration and residual retention of episomal plasmids, each iPSC line was migrated to feeder free conditions and grown beyond passage 15 on hESC-qualified Matrigel (1:100 dilution, BD Biosciences) coated plates in Essential 8 media (Life Technologies). DNA was extracted from feeder free cultures using DNeasy Blood and Tissue Kits (Qiagen). PCR was performed using 100 ng of genomic DNA, an annealing temperature of 72°C and 25 cycles using primers designed to amplify a region common to all episomal vectors used (supplementary table 2). Genomic DNA (100 ng) isolated from day 7 cultures, and 1 pg of each episomal vector were used as positive controls. PCR products were run on a 1% agarose gel and visualized using ethidium bromide.

2.4.9 Karyotyping

After 15 passages on MEF and hESC media, cells were migrated to 1:100 hESC Matrigel (BD Sciences) and maintained on Essential 8 media for upwards of 6 passages. Feeder-free adapted cells were sent to Cell Line Genetics Inc (Madison, WI) for karyotyping as described in [82].

2.4.10 Teratoma formation assays

in vivo developmental potential of the reprogrammed cell lines was examined. Monolayer iPSCs from three chimpanzee lines were grown on Matrigel (1:100) in E8 medium (Life Technologies) and collected by EDTA treatment (Life Technologies). Cells were counted and resuspended at a ratio of 1:1 cell volume to Matrigel and kept on ice until the injection. Six-week-old CB17.Cg-$Prkdc^{scid}Lyst^{bg\text{-}J}$/Crl immunodeficient male mice were obtained (Charles River Laboratories) and approximately one million iPSCs for each clone were injected into the testis-capsule. After five to eight weeks teratomas were isolated, weighed, measured, dissected, and fixed in 10% formalin. The specimens were embedded in paraffin, stained with hematoxylin and eosin, and analyzed by a histopathologist. All animal work was conducted under the approval of the Institutional Care and Use Committee of UCSD (Protocol# S09090). In addition, live feeder free iPSC cultures maintained in Essential 8 media on Matrigel iPSCs from C4955 (passage 15+7) were provided to Applied Stem Cell Inc. (Menlo Park, CA) for teratoma analysis as previously described [28].

2.4.11 Species-of-origin identity of teratoma samples

DNA was extracted from frozen teratoma tissue using DNeasy Blood and Tissue Kits (Qiagen). For teratomas derived from individual C4955, core sections were isolated from FFPE embedded teratomas tissue using a 3 mm dermal punch tool; DNA was extracted from core samples using a QIAamp DNA FFPE Tissue Kit (Qiagen). PCR was performed using

universal mitochondrial primers ([59] 2.2) amplifying cytochrome b (*Cytb*, chimpanzee reference sequence NC_ 001643:bp 14233-14598) or the 12S ribosomal gene (*12S*, NC _ 001643:bp 484-915) with 250-500 ng of genomic DNA as the starting template. Two-step PCR was conducted with an annealing temperature of 50°C for 1 minute and an extension step at 72°C for 4 minutes for a total of 30 cycles. DNA was purified using a Wizard SV gel and PCR Clean-up kit (Promega); dye terminator cycle sequencing was conducted by the University of Chicago Comprehensive Cancer Center using 60 ng of purified PCR template and 4μM of either the forward or reverse primer. Alignment to the chimpanzee, human (NC_ 012920) and mouse (NC_ 005089) reference sequences was accomplished using CLC Main Workbench 6.9 (Qiagen) and MUSCLE [38].

2.4.12 Directed differentiation of chimpanzee iPSCs to hepatocytes and cardiomyocytes

In order to demonstrate that chimpanzee iPSCs can be directly differentiated into other cell types, we differentiated C2 iPSC into hepatocytes and C7 into cardiomyocytes using the published protocols of [29] and [67] respectively, with the following modifications: In both cases we plated iPSCs at 0.35 x 10^6 cells/cm^2 in 0.44 mL/cm^2 and cultured them in Essential 8 media 24 hours prior to initiating all differentiations. To increase hepatocyte differentiation efficiency, 1μM of sodium butyrate was added during the first 24 hours of differentiation. After 24 days of differentiation, cells were immunostained as described above with a primary antibody for albumin (1:200, A6684, Sigma Aldrich; Supplementary Fig. 2.14). After 10 days of differentiation, differentiated C7 cultures were enriched for cardiomyoctes by culture in RPMI based media without glucose supplemented with 5 mM sodium DL-lactate for 10 days as described previously [22, 141]. After day 20 purified cardiomyocytes were cultured in media lacking glucose supplemented with 10 mM galactose [109]. After 25 days of cardiac differentiation, we characterized calcium flux in and out of iPSC-derived cardiomyocytes by

treating cultures with 5μM Fluo-4 AM (F-14217, Life Technologies) for 15 minutes, washing cultures once and imaging them with an AMG EVOS FL.

2.4.13 Microarray genotyping and PluriTest

RNA from passage $\geq$ 15 iPSCs was extracted using the QIAGEN RNeasy kit according to the manufacturer's instructions. Quality of the extracted RNA was assessed using an Agilent Bioanalyzer 2100 (RIN scores for all samples ranged from 9.9 to 10), and RNA was processed into biotinylated cRNA and hybridized to the HT12v4 array using standard Illumina reagents as directed by manufacturer. Arrays were scanned using an Illumina HiScan, and data processed using Illumina's GenomeStudio software. Using these data, we carried out PluriTest as previously described [87]. Additionally, we mapped all detected HT12v4 probe sequences (n = 46,297) to the chimpanzee (panTro3) genome using BWA 0.6.3 [65]. Probes that mapped to a single genomic location with no mismatches were retained (n = 21, 320, 46.2% of all probes) for the analysis that was restricted only to the chimpanzee lines. When we considered data from human and chimpanzee iPSCs together, without excluding probes based on sequence matches to the chimpanzee genome, all chimpanzee lines in the panel had pluripotency scores slightly below the pluripotency threshold (Supplementary Fig. 2.17, lighter points). However, low pluripotency scores could stem from differences in our ability to estimate gene expression levels in the chimpanzee compared to the human due to attenuated hybridization caused by sequence divergence [44]. Indeed, when we subset the array to retain only those detected probes that map to the chimpanzee genome with no ambiguity or mismatches, all chimpanzee lines have pluripotency scores greater than the pluripotency threshold value of 20 (Supplementary Fig. 2.17, darker points).

2.4.14 RNA sequencing and differential expression testing between iPSCs

50 bp single-end RNA sequencing libraries were generated from RNA extracted from 7 chimpanzee and 7 human iPSC lines using the Illumina TruSeq kit as directed by the manufacturer (San Diego, CA), as well as from their precursor fibroblast or LCL cell lines. All iPSC samples were multiplexed and sequenced on four lanes of an Illumina HiSeq 2500; while the precursor cell lines were multiplexed and sequenced on six lanes of the same sequencer. We generated a minimum of 28,010,126 raw reads per sample (supplementary table 9), and confirmed the raw data were of high quality using FastQC (available online at http://www.bioinformatics.babraham.ac.uk/projects/ fastqc/). We mapped raw reads to the chimpanzee (panTro3) or human (hg19) genome as appropriate using TopHat 2.0.8 [143], allowing for a maximum of 2 mismatches in each read. Due to the relatively poor annotation of the chimpanzee genome and to prevent biases in expression level estimates due to differences in mRNA transcript size and genetic divergence between the two species, we limited the analysis to reads that mapped to a list of orthologous metaexons across 30,030 Ensembl genes drawn from hg19 and panTro3, as in [13]. Following mapping, gene level read counts were generated using *featureCounts* 1.4.4 as implemented in Subread [68]. Due to mapping biases between human and chimpanzee ribosomal proteins and pseudogenes, we removed all genes associated with the Gene Ontology Cellular Compartment category 'ribosome' (GO:0005840, n = 141) and all annotated pseudogenes in Ensembl release 65 (n = 3170, December 2011, the oldest available archival version of Ensembl) from the data at this point. We considered two normalization approaches in our analysis. In one instance, we examined only RNA-sequencing data from chimpanzee and human iPSCs, and retained 12,171 genes with at least 4 observations in one of the two species of $\log_2$ CPM > 1. CPM were then loess normalized by species within individuals with *voom* [62]. As the orthologous genes are not constrained to be the same length in both species, we computed RPKM for each gene before carrying out any inter-species comparisons. We then used the R/Bioconductor pack-

age *limma* 3.20.3 [128] to test for differential expression in our RNA-seq data, with a model that included only a species effect. Finally, we tested for an enrichment of GO categories amongst DE genes using the R package *topGO* 2.16.0 [2]. These normalized values were used only to identify genes DE between iPSCs of the two species. For the dataset containing RNA-sequencing data from iPSCs and their precursors, we again only retained 13,147 genes with at least 4 observations in one of the four groups (chimpanzee iPSCs, chimpanzee precursors, human iPSCs or human precursors) of $\log_2$ CPM > 1. Gene counts were then loess normalized within individuals by tissue, after correcting for the lack of independence within different tissues from the same individual, through the function corfit. As above, we then computed species-specific RPKM values, and used *limma* and *topGO* to test for differential expression and GO category enrichment, respectively. In this instance, we used a model design with 6 parameters for the main effect (chimpanzee iPSC, human LCL-derived iPSC, human fibroblast-derived iPSC, chimpanzee fibroblast, human LCL and human fibroblast) and no additional covariates. To confirm that our conclusions are robust with respect to the choice of normalization procedure, in both cases, we also tried a variety of other normalization schemes, including correcting for %GC content as in [112], none of which had a substantial effect on the final results (2.6). Finally, we built neighbor joining trees using Manhattan distances calculated from RPKM values at all 13,147genes using the nj function in the R library *ape* [101]. All analyses were performed at a false discovery rate [10] threshold of 1% unless otherwise noted, using R 3.1.0 and Bioconductor 2.14 [43].

2.4.15 DNA Methylation arrays

To analyze DNA methylation, we extracted DNA from all chimpanzee and human iPSC lines described above, as well as from the source fibroblast or LCLs. In all cases, 1000 ng of genomic DNA were bisulphite-converted and hybridized to the Infinium HumanMethylation450 BeadChip at the University of Chicago Functional Genomics facility as directed by

the manufacturer. Since the probes on the array were designed using the human reference genome, we followed the approach described in [47] to compare humans and chimpanzees. We retained those probes that had either a perfect match to the chimpanzee reference genome, or had 1 or 2 mismatches in the first 45 bp but no mismatches in the 3' 5 bp closest to the CpG site being assayed. We also removed all probes that contained human SNPs (MAF ≥ 0.05) or chimpanzee SNPs (MAF ≥ 0.15) within the last 5 bp of their binding site closest to the CpG being assayed. Within each individual, probes with a detection $P > 0.01$ were excluded. This resulted in the retention of 335,307 autosomal probes, and an additional 8,210 X chromosome probes, which we normalized and analyzed separately by sex. In all cases we performed a two-color channel signal adjustment, quantile normalization and β-value recalculation as implemented in the lumi package [37]. Because the HumanMethylation450 BeadChip contains two assay types which utilize different probe designs, we performed a BMIQ (beta mixture quantile method) normalization [140] on the quantile-normalized autosomal data set. We did not perform this step on the X chromosome data, due to its methylation patterns. We built neighbor joining trees using Manhattan distances at all 335,307 probes using the nj function as above. In order to identify DM probes we used an identical approach to that described above for the identification of DE genes. First, we identified probes that were DM between the iPSCs of both species using *limma* by using a reduced data set and model containing only data from the iPSCs themselves. Then, we fit a linear model to the data using limma with 6 parameters corresponding to the 6 tissue/species combinations in the data, classifying probes as DM at an FDR of 1%. As with the expression data, the reduced model has more power to identify DM probes between the two iPSC groups than the full model; however, there is great concordance between the two sets of results (Supplementary Fig. 2.31). We excluded all probes with mean βinter-group differences < 0.1 in order to group DM probes into DMRs, which we define as 2 or more DM probes separated by $<$ 1kb, with the additional requirement that the effect be in the same direction in all DM probes

within the region. Finally, to examine the content of these DMRs, we used annotation files for the HumanMethylation450 Bead Chip provided by the manufacturer and discarded all DMRs associated with either multiple or no genes. We tested for enrichment of GO BP categories amongst the genes contained in the DMRs by using the R package topGO 2.16.0 [2], using as a background set all genes in which it is theoretically possible to detect DMRs.

2.4.16 H3K27ac and H3K27me3 ChIP-seq data

ChIP-seq assays were performed as previously described [121], with slight modifications. Specifically, approximately 60 million iPSCs from three chimpanzee individuals (C2, C5 and C7) were cross-linked with 1% formaldehyde for 10 minutes. Cells were lysed and chromatin sheared with a Covaris S2 (settings: 4 minutes, duty cycle 10%, 5 intensity, 200 cycles per burst in 4 6x16mm tubes per individual). H3K27ac- and H3K27me3-enriched regions were isolated using 5 ug of either H3K27ac antibody (ab4729, Abcam, Cambridge, MA, USA) or H3K27me3 antibody (07-449, Millipore, Billerica, MA, USA). ChIP and input DNA from each individual were end-repaired, A-tailed and ligated to Illumina Truseq sequencing adapters before 18 cycles of PCR amplification. 200-300 bp DNA fragments were selected for sequencing. Input libraries were multiplexed and sequenced on one lane of an Illumina HiSeq2500 using the rapid run mode, ChIP libraries were multiplexed and sequenced on three lanes of an Illumina HiSeq2500 using the rapid run mode. For comparison purposes, we downloaded ChIP input, H3K27ac and H3K27me3 data from 3 human iPSC lines (iPS 6.9, iPS-18a, and iPS.20b, all of them release 5) generated by the Roadmap Epigenomics Consortium [114] from the NIH GEO database (2.6. Human and chimpanzee samples were mapped to either hg19 or panTro3 using BWA 0.7.9 [65]; reads that mapped outside chromosomes 1-22 + X were discarded, as were reads that did not map uniquely to a single genomic region with less than 2 mismatches, or reads that were marked by Picard (http://picard.sourceforge.net) as originating from PCR duplicates. After mapping and fil-

tering, we used MACS 1.4.4 [171] and RSEG 0.4.4 [134] to identify peaks in the H3K27ac and H3K27me3 data respectively. Our analyses in this section follow those of [174]. Briefly, for MACS, we specified an initial p-value threshold of H3K27ac, 0.001, and used each line's ChIP input file for comparison. For RSEG, we used the 'rseg-diff' function to compare H3K27me3 enrichment against each individual's ChIP input file, with the recommended 20 maximum iterations for hidden Markov model training. We then filtered enriched regions or peaks identified by either program by retaining only those that overlapped a previously defined set of 200 bp orthologous windows [174], where at least 80% of bases are mappable across all three species using liftOver. We define mappability as the ability of each 20bp kmer beginning in that window to be uniquely mapped to the genome. To ensure that sequence divergence did not confound our analyses, we mapped each identified region or peak in humans to the chimpanzee genome, and vice versa, using liftOver, and excluded regions and peaks where 80% or greater of bases in the enriched peaks or regions failed to align to the other genome. To further minimise the number of false positive results in our inter-species comparison (due to incomplete power), we applied a two-step cutoff [23] to the list of enriched regions and peaks. For H3K27ac, we retained all peaks that were identified with a first, stringent cutoff of FDR $< 5\%$ in one species and a, second, relaxed cutoff of FDR $< 15\%$ in the other, as in [174]. Because RSEG does not report FDR values for enriched regions, we used each region's domain score, which is the sum of the posterior scores of all bins within the domain, and set a first, stringent cutoff of 20 in one species, and a second, relaxed threshold demanding only that the region be classified as 'enriched' by RSEG, without a specific score requirement. Having done this, we integrated data from multiple peaks (when present) to generate a gene-level metric of ChIP signal in each individual. Specifically, we computed an enrichment score for each histone mark in each individual in a set of previously defined 26,115 orthologous TSSs [174] by dividing RPKM values at each TSS at gene i for either mark minus RPKM values in TSS at gene i for ChIP input, all of it over

the genome-wide average RPKM for either mark minus the genome-wide average RPKM for ChIP input. Given the way in which we have defined this enrichment score, a score > 0 indicates those genes where we detected more histone mark reads than input reads, while a score > 1 indicates a gene with an excess of histone mark reads greater than what we would expect given the genome-wide distribution. Because 8 of the 22 genes in the list of pluripotency master genes used to generate Fig. 2.5D do not have clearly defined orthologous TSSs, we also examined whether MACS identified peaks in the 2kb +/- TSS for all 22 genes and their orthologous position in the chimpanzee genome, identified solely through liftOver - that is, without taking into account whether there is evidence for a TSS at that position in the chimpanzee genome. To generate Supplementary Fig. 2.22, we asked how many of the 22 genes had at least 1 peak at an FDR $< 5\%$ in at least 1 individual in either species, regardless of orthology and sequence conservation. We note that since different labs produced the human and chimpanzee data, we expect a considerable technical batch effect to be completely confounded with species annotation. Given this study design, we expect the technical batch effect to result in the appearance of inter-species differences; yet, our goal is to demonstrate similarity across species. Thus, our conclusions (of high overlap across species), are conservative with respect to the technical batch effect.

2.4.17 REX1 expression and function

To examine the possible consequences of reduced expression of *REX1* in chimpanzee iPSCs, we retrieved genes that responded to a *REX1* knockdown in mESCs from supplementary tables 2 and 3 of [123] and converted Affymetrix MG 430 2.0 probe IDs to ENSM and ultimately to orthologous ENSG identifiers using Biomart release 66 (to control for deprecated identifiers). Because [80, 123, 132] have highlighted *REX1's* function in controlling cell cycle progression, glycolysis and cellular differentiation, we additionally retrieved genes associated with these terms to generate Fig. 2.6C as follows: The core set of pluripotency TFs are those

described by [98] and [91]. Cell cycle and glycolysis categories contain all genes associated with GO BP:0007049 and BP:0006096 respectively, whereas cell fate contains genes associated with any GO term that contains the words "ectoderm", "mesoderm" or "endoderm". We also examined individual examples of cell fate differentiation: CNS development genes are associated with BP:0007417 or any of its offspring; cardiovascular system development genes are associated with BP:0072358 or any of its offspring; hepatobiliary system development genes are associated with BP:0055123 or any of its offspring. Confidence intervals around the null hypothesis were generated independently for each category from 100,000 permutations in R. Finally, to compare our data with a previously published set of chimpanzee and bonobo iPSC lines [76], we downloaded fastq files from GEO (Series GSE47626) and mapped only the first mate from all reads using the same approach as above, but allowing 4 mismatches in the entire 100bp read. We normalized expression estimates jointly with our own cell lines to generate Supplementary Fig. 2.24, and used all data points from a given species, irrespective of origin, to generate boxplots. To generate Supplementary Fig. 2.25, we took Illumina HT12v4 array data from 73 human iPSC lines generated in house and calculated Pluritest pluripotency and novelty scores as above, using the full probe set. Independently, we normalized and background-corrected the raw array intensities using the lumiExpresso function in *lumi* and extracted expression values for all 73 human iPSC lines at the single array probe associated with *REX1*. In order to examine *REX1* methylation levels in other human PSCs, we obtained Infinium HumanMethylation450 BeadChip for the lines reported in [175] from the authors, and normalized it jointly with our own data as described above. We then extracted normalized methylation βlevels at the 13 probes that map to *REX1* in both chimpanzees and humans to generate Supplementary Fig. 2.26.

2.4.18 Other indicators of genomic stability

Finally, we assessed two broad indicators of stability in our chimpanzee lines. All iPSC lines derived from female chimpanzees, and 3 of 4 lines derived from human females, show strong evidence for elevated expression of *XIST* relative to male lines (FDR-adjusted $P = 0.0010$; Supplementary Fig. 2.32)) and maintenance of X-chromosome inactivation during pluripotency. X-chromosome methylation patterns in females corroborate these observations, with the majority of probes mapping to the X-chromosome in our data being either hemimethylated ($0.2 < \beta < 0.8$) or hypermethylated ($\beta \geq 0.8$) in females but not in males (Supplementary Fig. 2.33). We also used a list of 168 imprinted probes from [75] to check for maintenance of genomic imprinting after reprogramming. We find that the majority of imprinted loci remain hemimethylated following reprogramming in both human and chimpanzee iPSC lines (Supplementary Fig. 2.34). However, we identify two sets of probes that are consistently hypermethylated in pluripotent lines but were hemimethylated in their precursor cells. The first cluster contains 5 probes that are hypermethylated across both chimpanzee and human iPSCs; these probes are associated with the genes *KCNK9*,*ANKRD11* and *MKRN3*. The second cluster is comprised of 21 probes that are hypermethylated in all human iPSCs but only 2 chimpanzee iPSCs in our data, and is associated with the gene *PEG3-ZIM2*, which has been previously shown to be abnormally methylated in both hESCs and hIPSCs [74].

2.4.19 Data Access:

All novel RNA-sequencing, DNA methylation and ChIP-seq data are available at the GEO under SuperSeries number GSE61343. Additionally, a table with P-values for all hypothesis testing performed using the methylation data (by probe) is available on the Gilad lab website (http://giladlab.uchicago.edu/ Data.html). All chimpanzee iPSC lines described in this publication are available fully and without restrictions to other investigators upon request to the corresponding authors.

2.5 Supplementary Information

2.5.1 Supplementary Figures

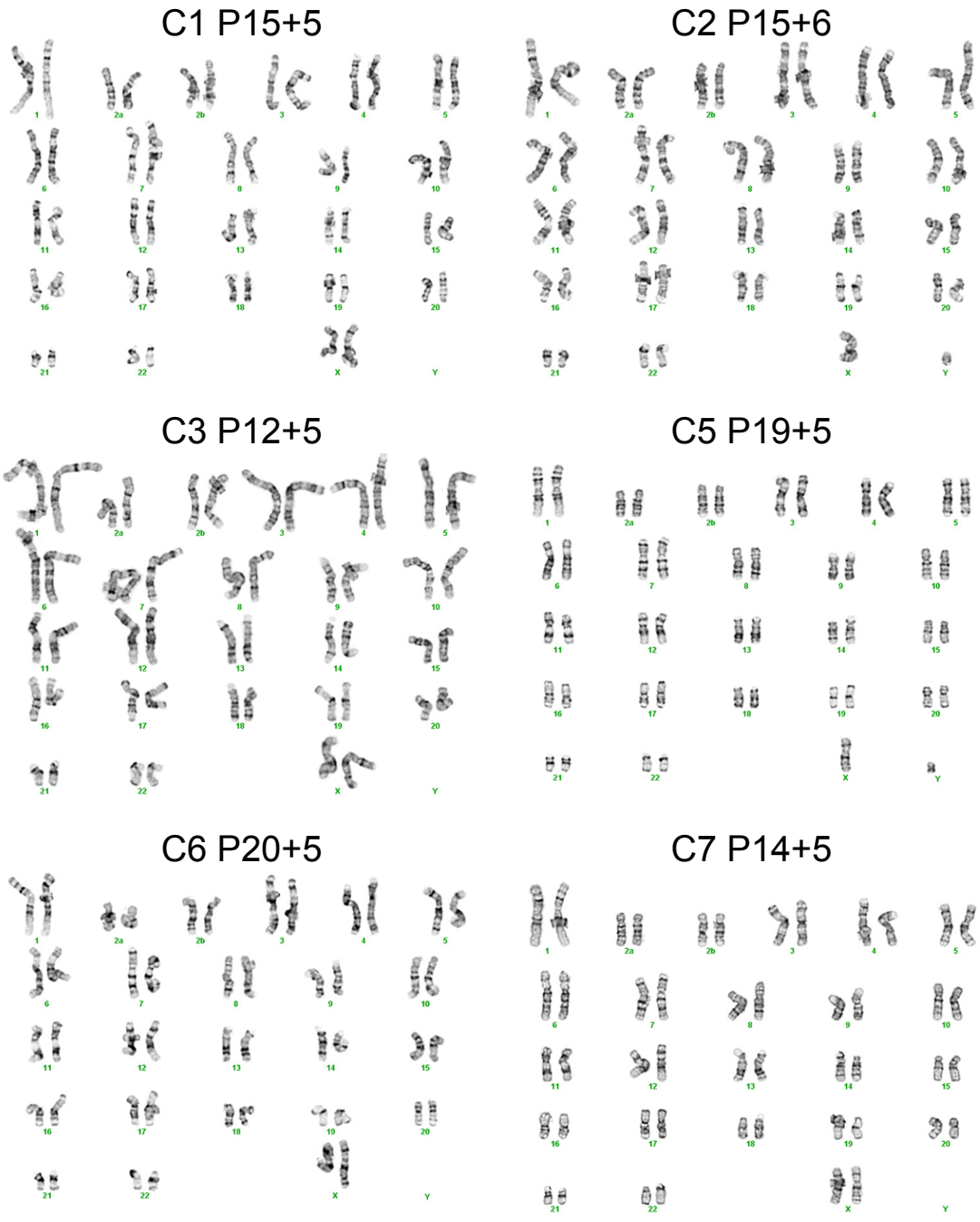


Figure 2.8: **Karyotypes for the 6 chimpanzee iPSC lines not shown in main text.** Figures, generated after >15 passages in culture. Passage number for each line represents passages on MEF feeders plus additional passages on Matrigel.

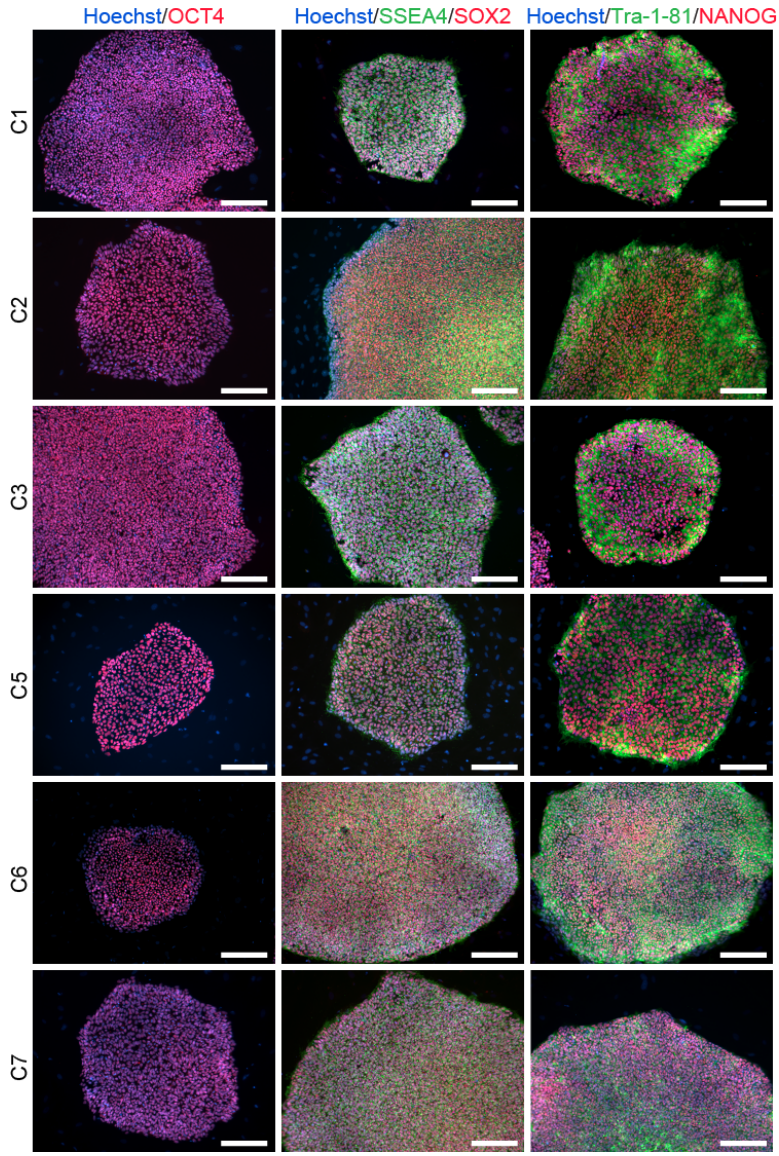


Figure 2.9: **: ICC staining of the 6 chimpanzee iPSC lines not shown in main text.** Figures with antibodies for pluripotency markers as indicated. Scale bar: 200μM.

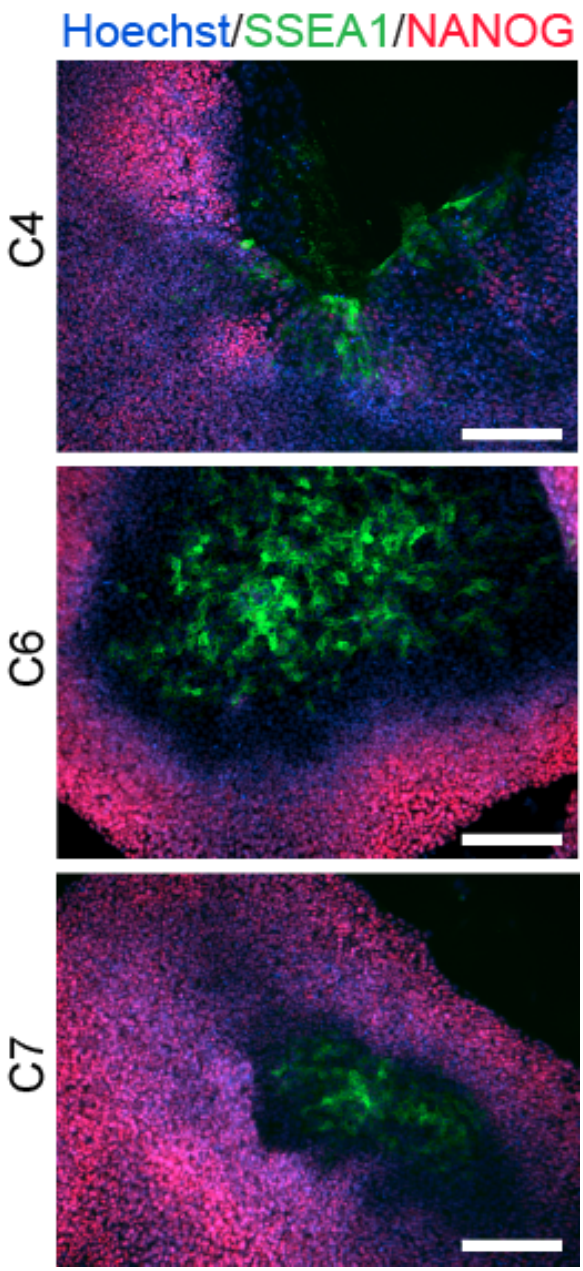


Figure 2.10: **ICC staining showing SSEA1 expression in chimpanzee iPSC culture plates, clearly distinct from NANOG expression.**

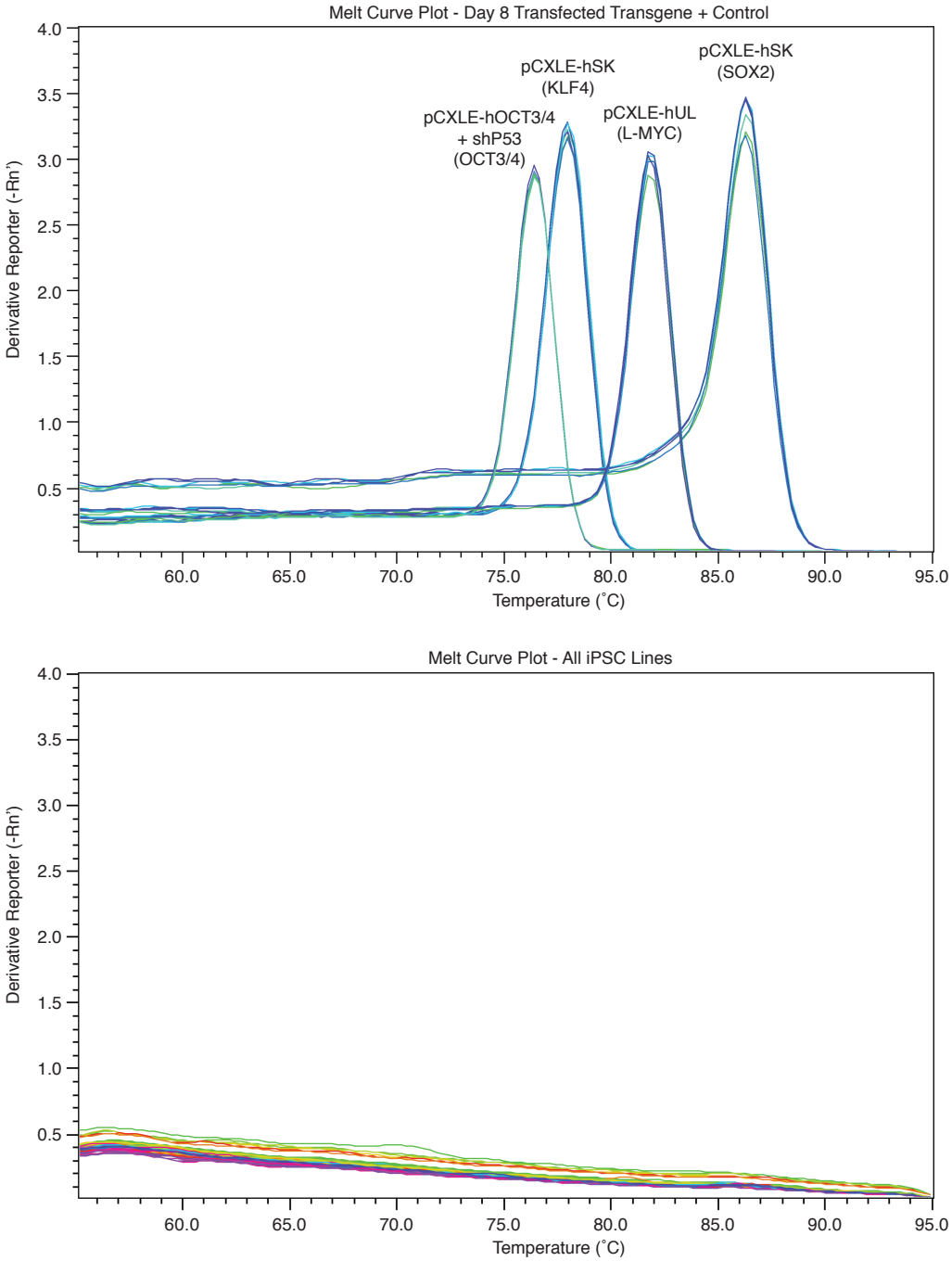


Figure 2.11: **Melt curves showing a lack of exogenous reprogramming gene expression in episomally reprogrammed chimpanzee iPSCs after > 10 passages.**

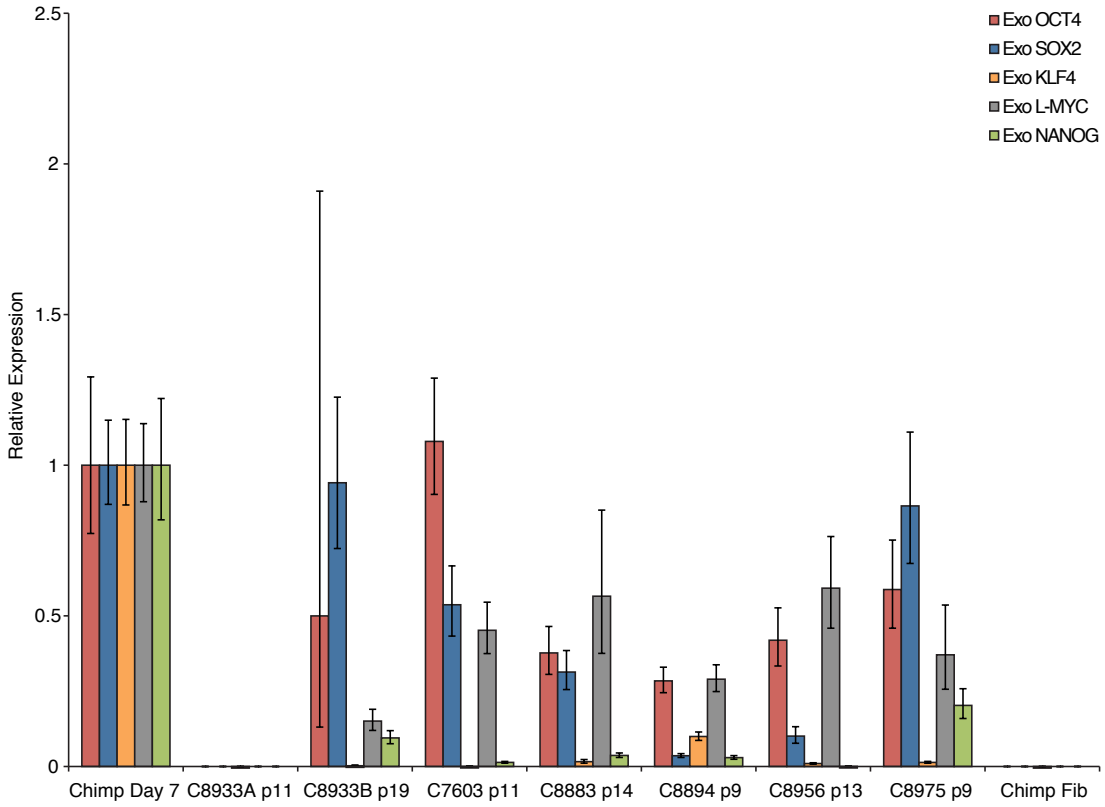


Figure 2.12: **Exogenous gene expression in retrovirally reprogrammed chimpanzee iPSCs after various passages. All values are relative to expression in a day-7-post-transfection chimpanzee fibroblast.**

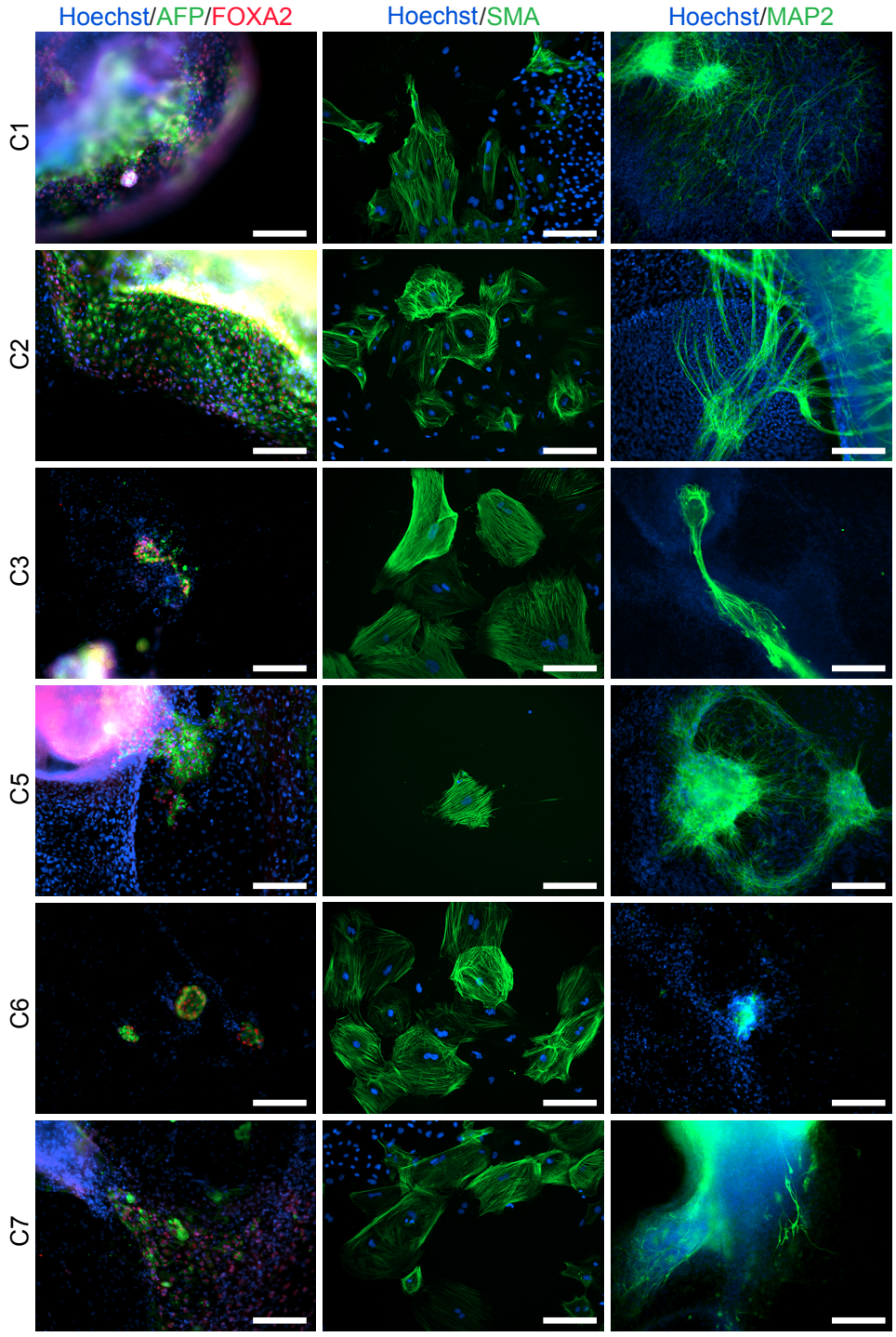


Figure 2.13: **ICC staining of differentiated embryoid bodies derived from the 6 chimpanzee iPSC lines not shown in main text Figures, with antibodies indicating the three germ layers as indicated.** Scale bar: 200μM.

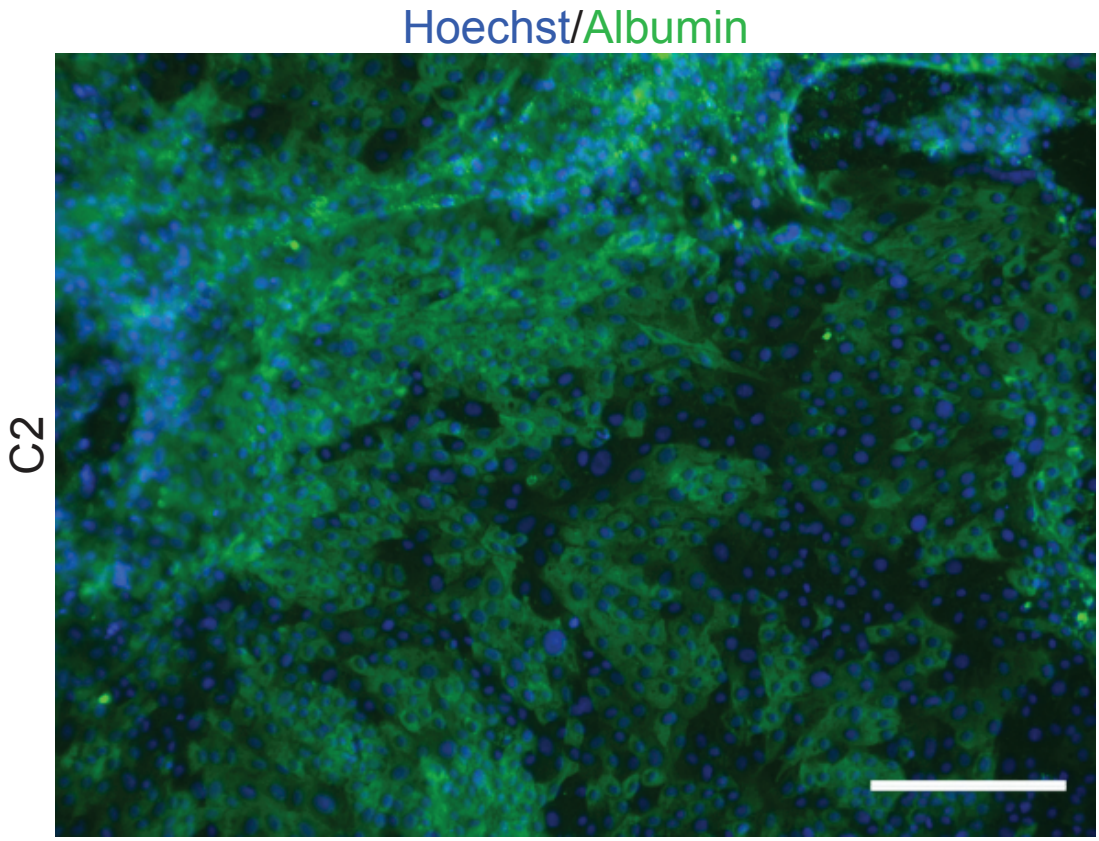


Figure 2.14: **ICC staining of directly differentiated hepatocytes from line C2, with antibodies as indicated.** Scale bar: 200μM.

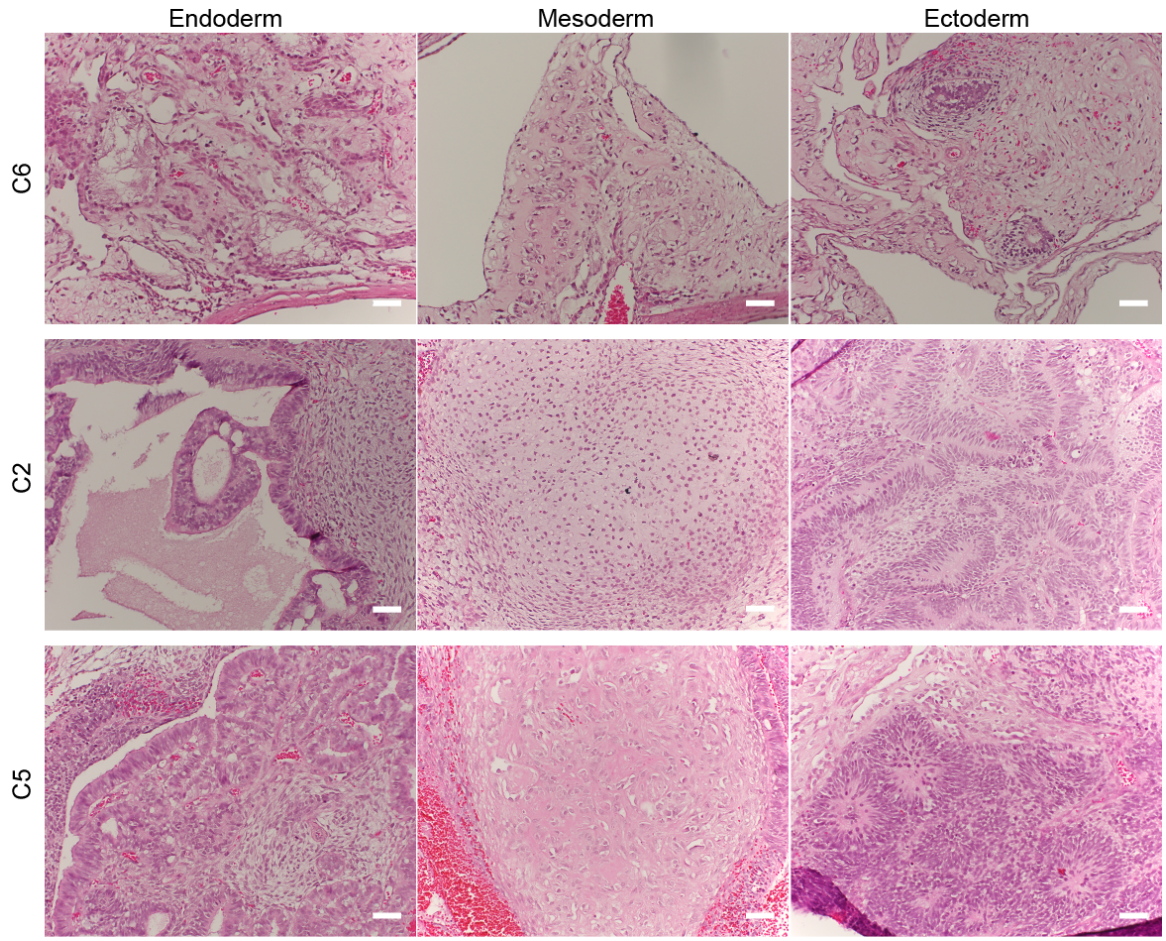


Figure 2.15: **Histological staining of teratomas derived from three additional chimpanzee iPSC lines, showing generation of tissues from all three germ layers.** Scale bar: 500μM.

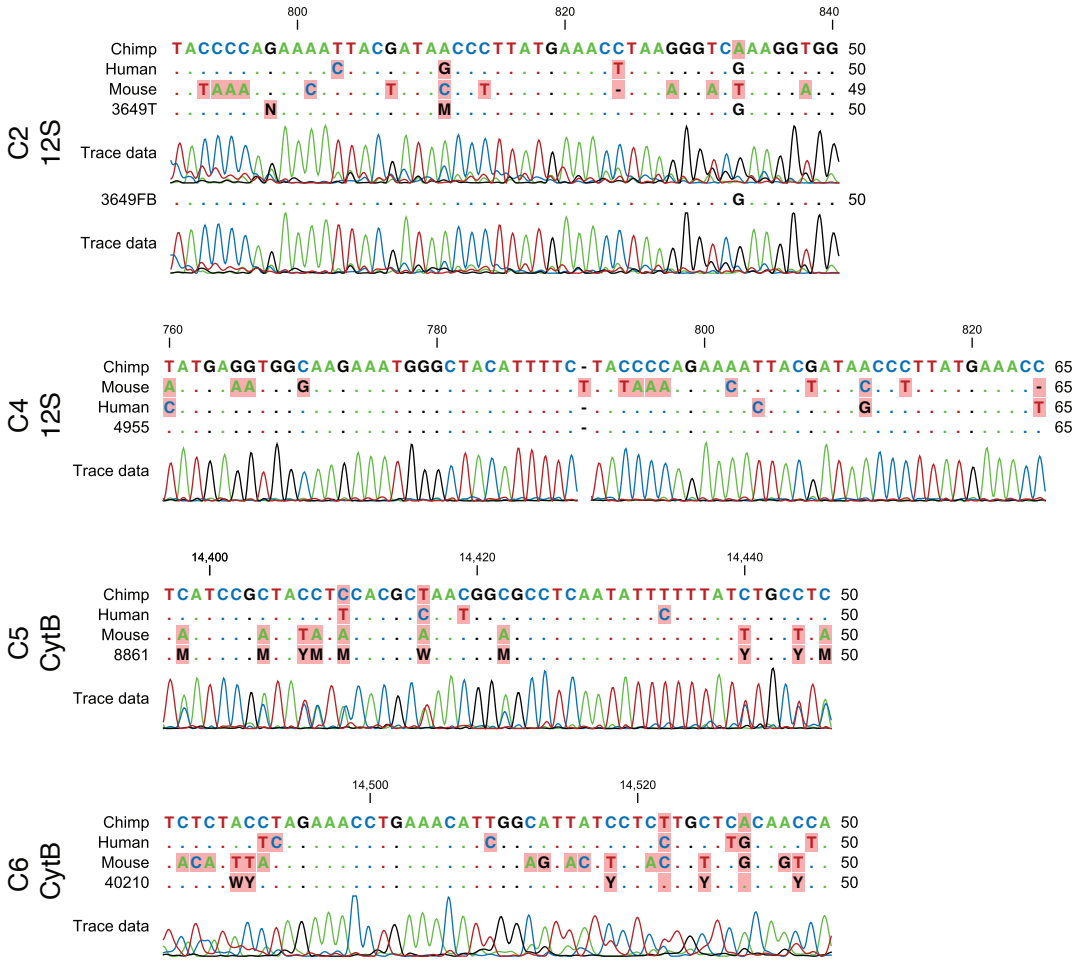


Figure 2.16: **Sequencing traces from teratomas generated from chimpanzee iPSC lines for the mitochondrial genes *12S* (C3649, C4955) and *cytb* (C8861, C40210).** All traces show clear evidence of the presence of chimpanzee tissue in the teratoma.

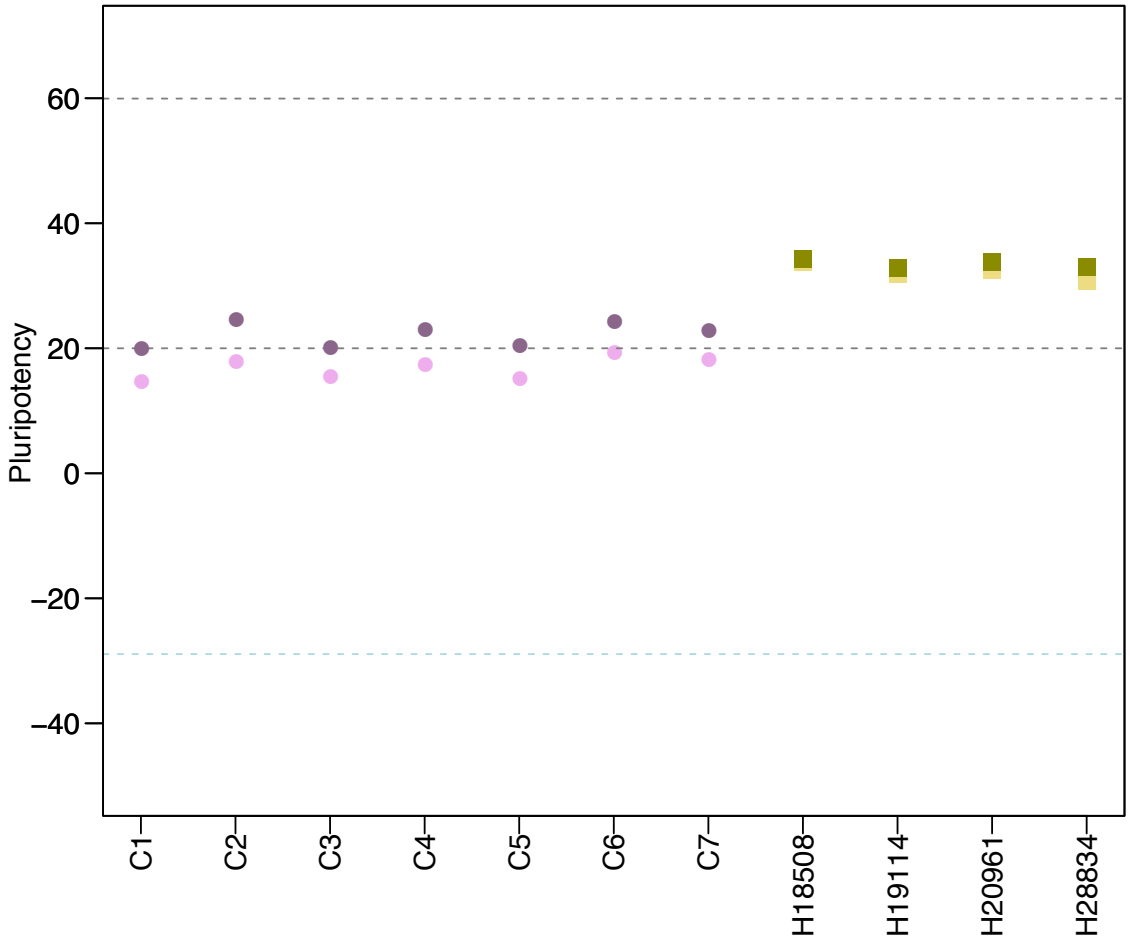


Figure 2.17: **The effects of probe sub-setting in PluriTest pluripotency score calculations.** Lighter shades indicate pluripotency scores before the removal of probes not mapping to the chimpanzee genome, darker shades indicate pluripotency after probe removal. Purple circles denote chimpanzees; yellow squares, humans.

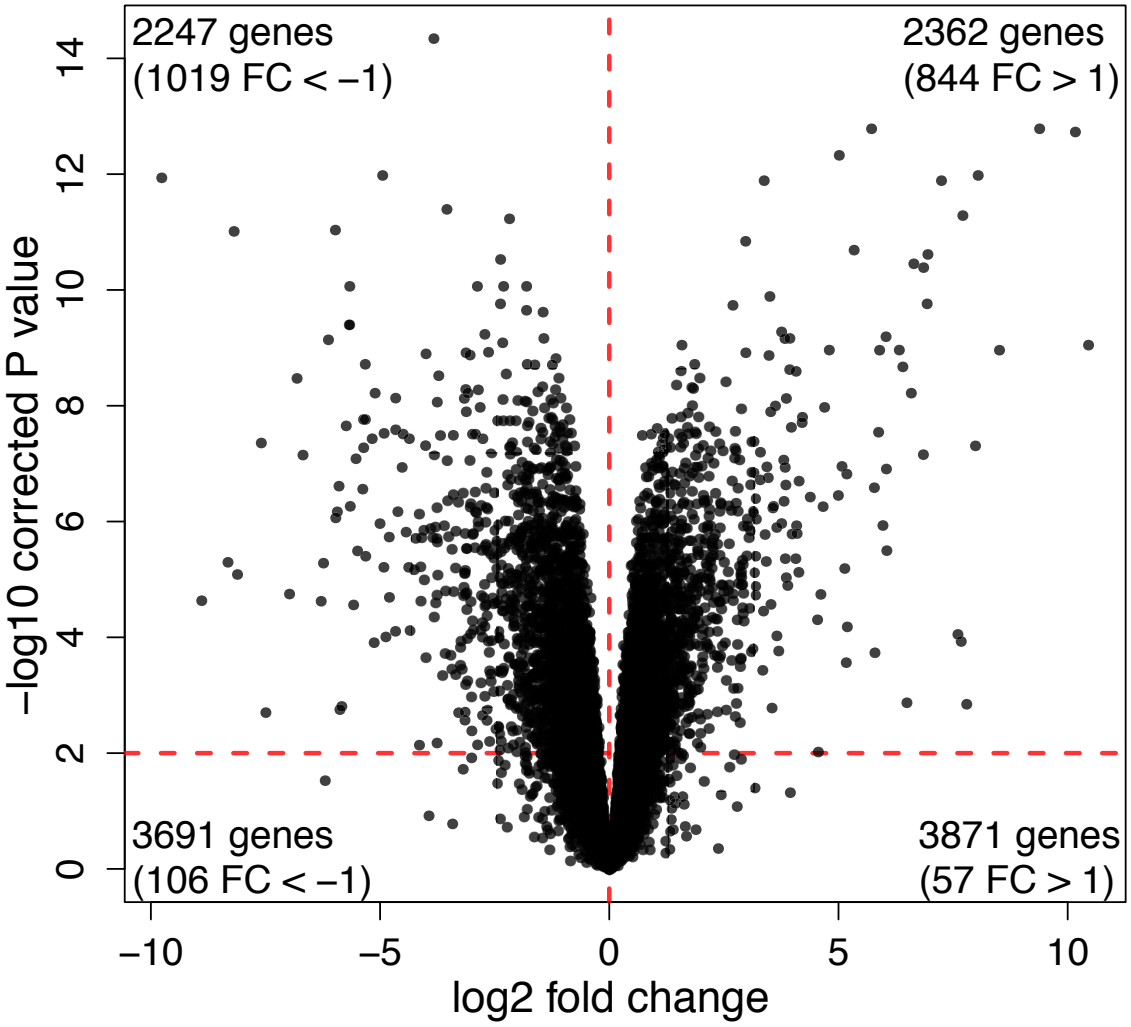


Figure 2.18: **Volcano plot showing the distribution of differentially expressed genes between iPSCs of chimpanzee and human origin.**

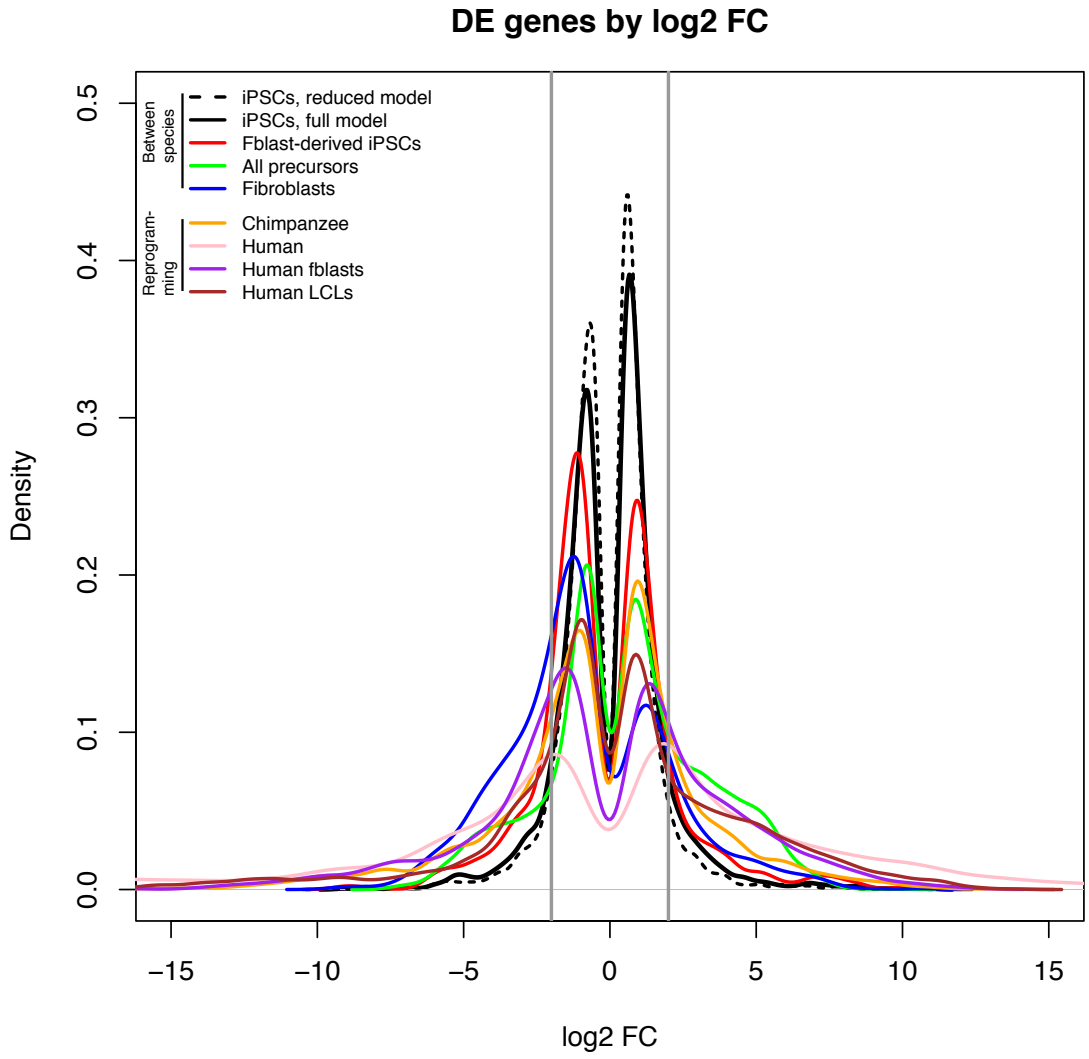


Figure 2.19: **Density plots of $\log_2$ FC change values amongst DE genes for the main comparisons presented in the text.** The area bounded by the grey lines represents $\log_2$ FC changes with an absolute magnitude < 2.

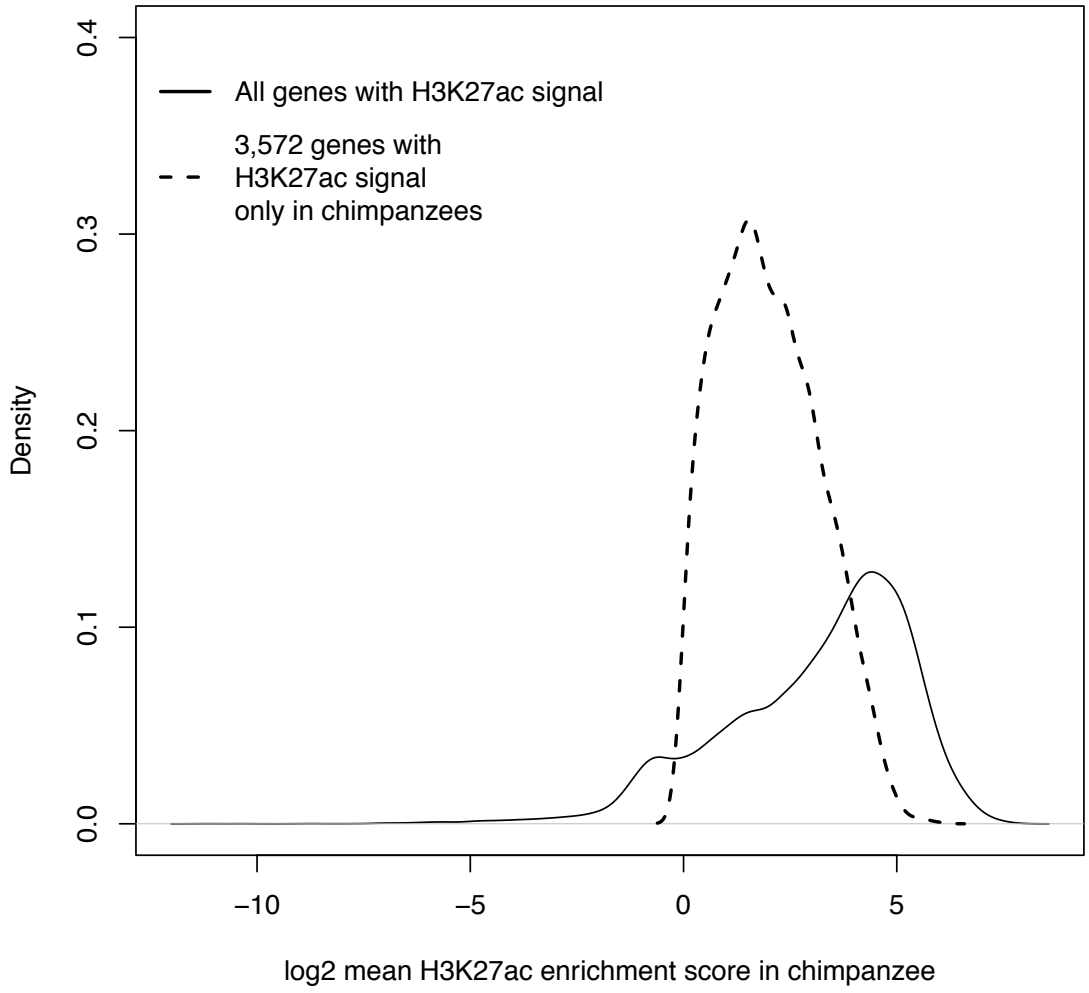


Figure 2.20: **Density plots of H3K27ac enrichment scores in the entire data set and at 3,572 genes enriched only in chimpanzee iPSCs.**

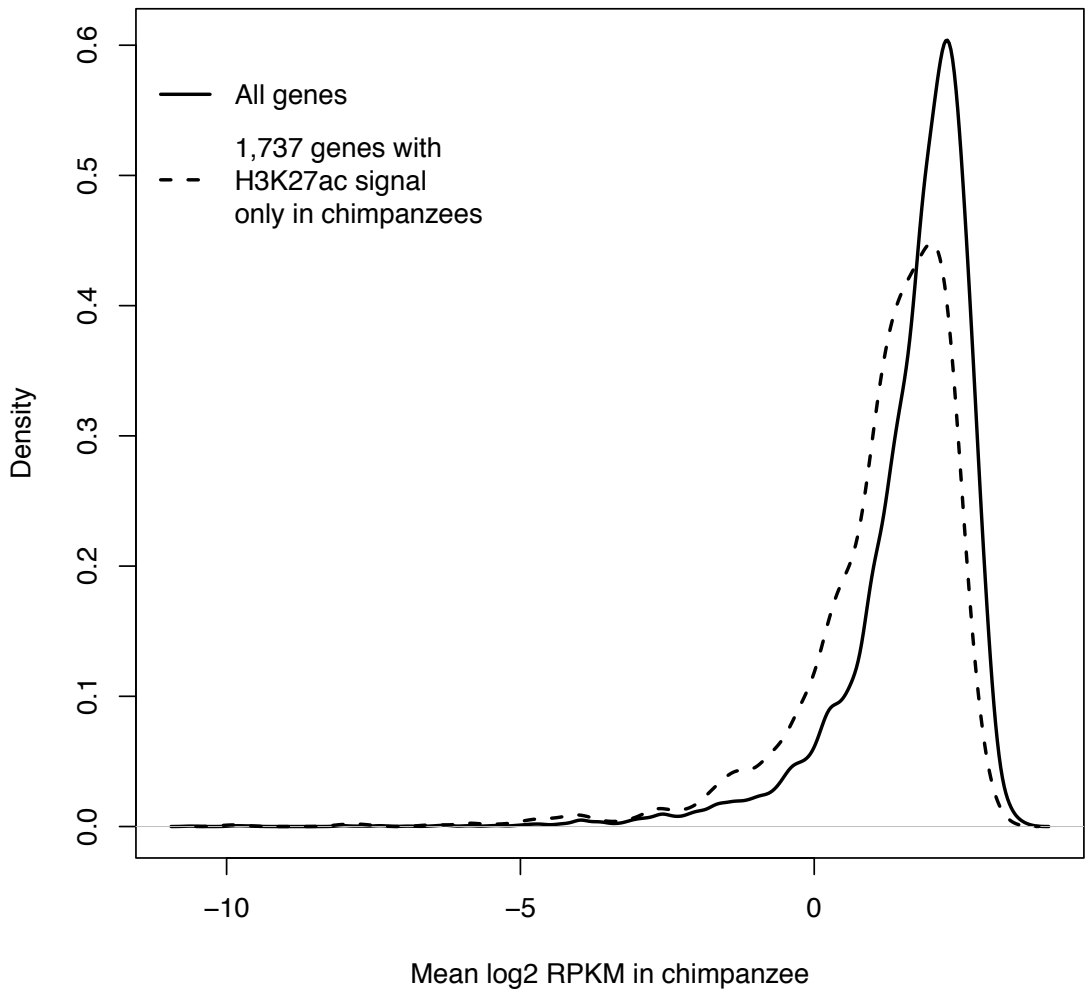


Figure 2.21: **Density plots of mean RPKM in chimpanzee iPSCs in all 12,171 genes with expression data and in the subset of 1,737 genes with expression data and H3K27ac signal enrichment solely in chimpanzee iPSCs.**

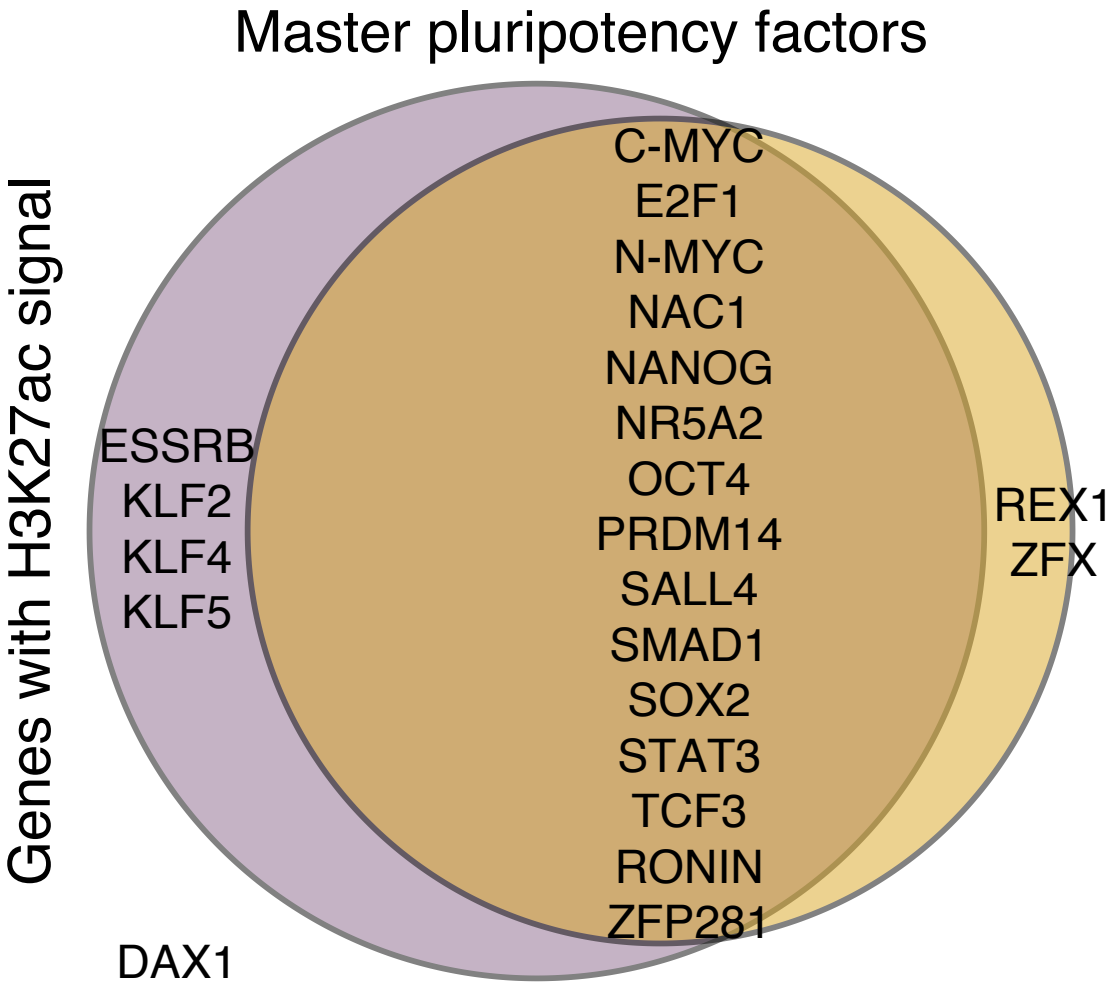


Figure 2.22: **H3K27ac peaks observed in at least 1 chimpanzee or human iPSC, as identified by MACS at 22 known pluripotency master regulators.** In the case of all three genes that differ between this Figure and Figure 5d - *KLF5*, *NR5AD* and *SMAD1* - processed enrichment signal after accounting for orthology is weak, and falls very close to our normalized enrichment score threshold of 1, explaining the difference between the two.

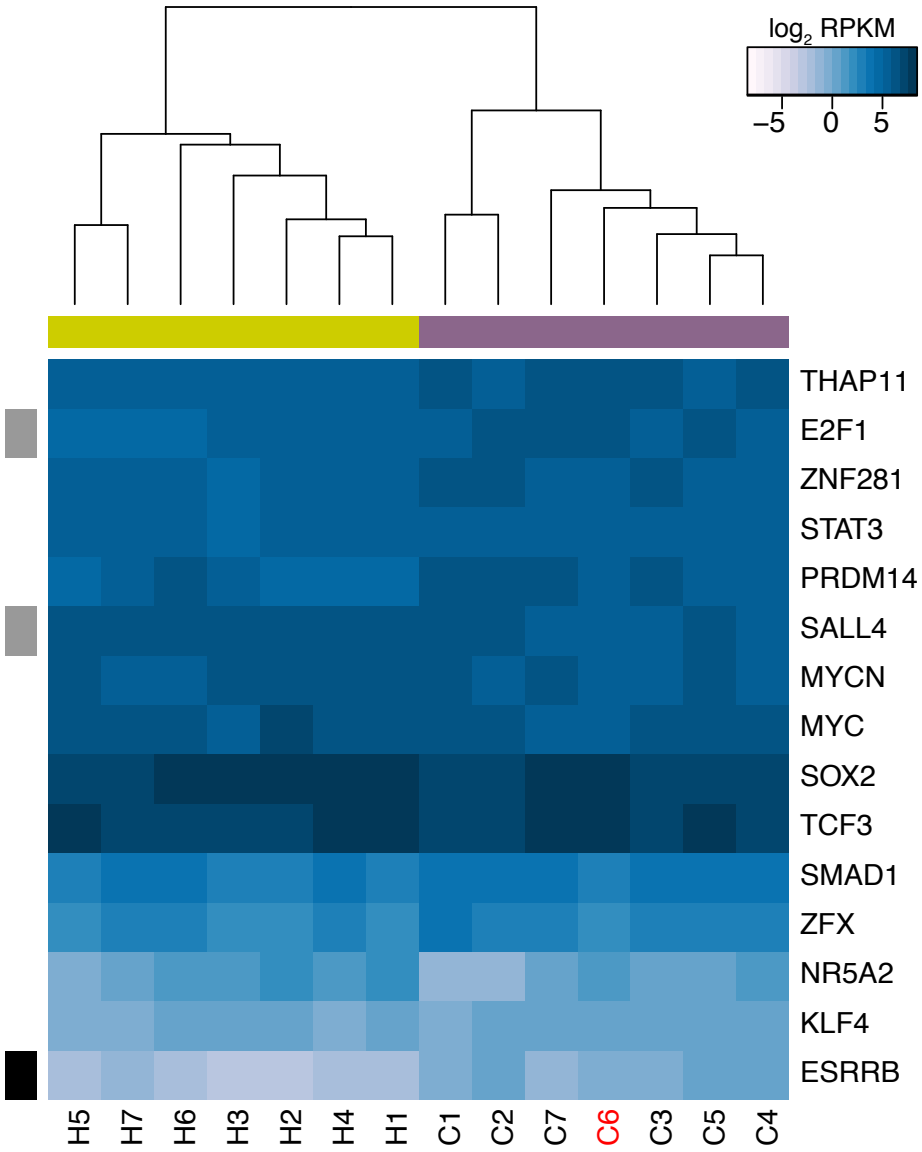


Figure 2.23: **Expression values of 15 core pluripotency transcription factors in all human and chimpanzee iPSC lines.**The data used to generate this Figure are identical to those used to generate Figure 5a except that expression levels of *REX1* are not included in the calculation. *REX1*-expressing chimpanzee iPSC line C6 is colored red, significant interspecies differences are indicated along the left-hand side, and purple boxes indicate chimpanzee lines, yellow boxes indicate human lines.

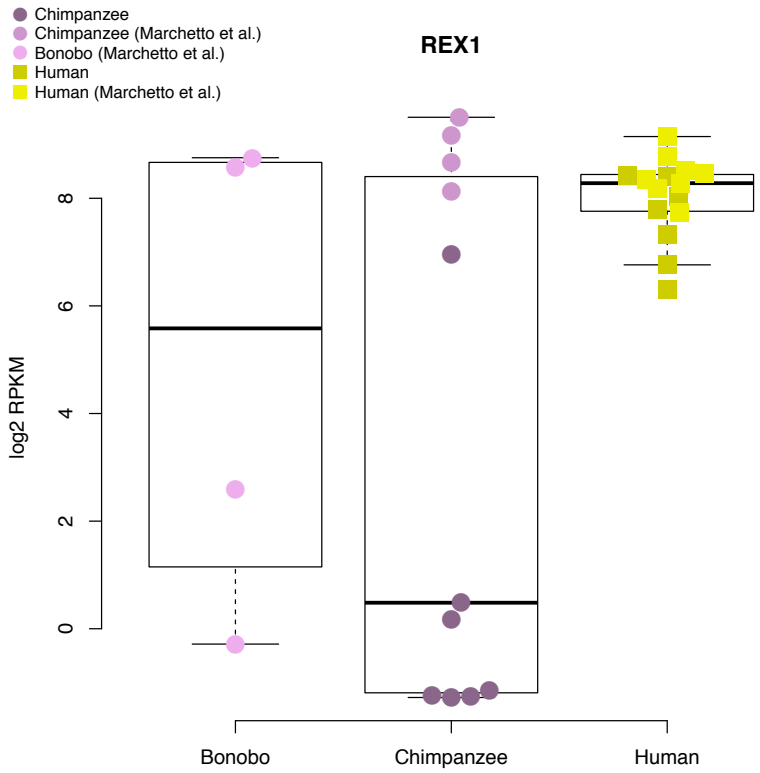


Figure 2.24: **Expression levels of *REX1* in human, chimpanzee and bonobo iPSC lines generated in this study and in [76].** Data for this Figure were jointly normalized.

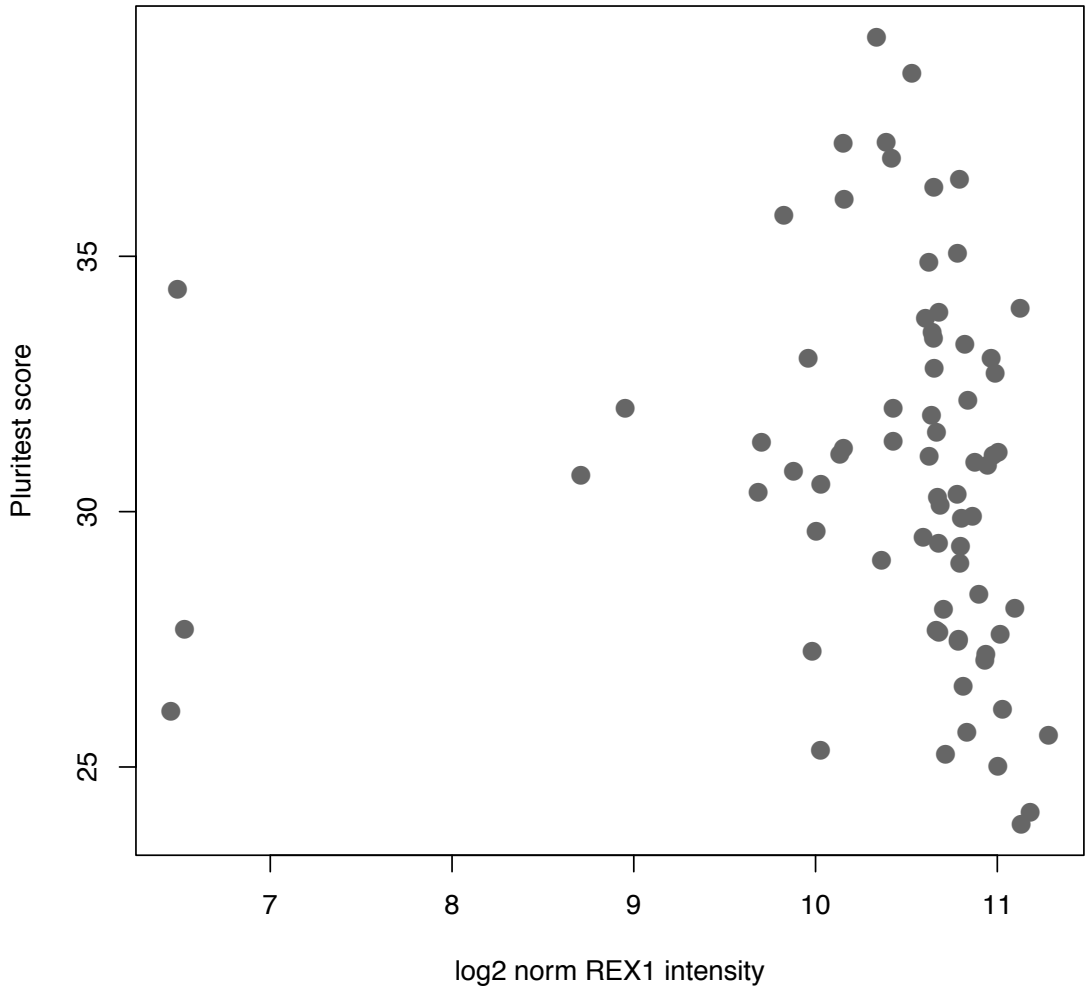


Figure 2.25: **Plot of PluriTest pluripotency scores versus normalized *REX1* intensity in 73 human iPSC lines derived in-house.**

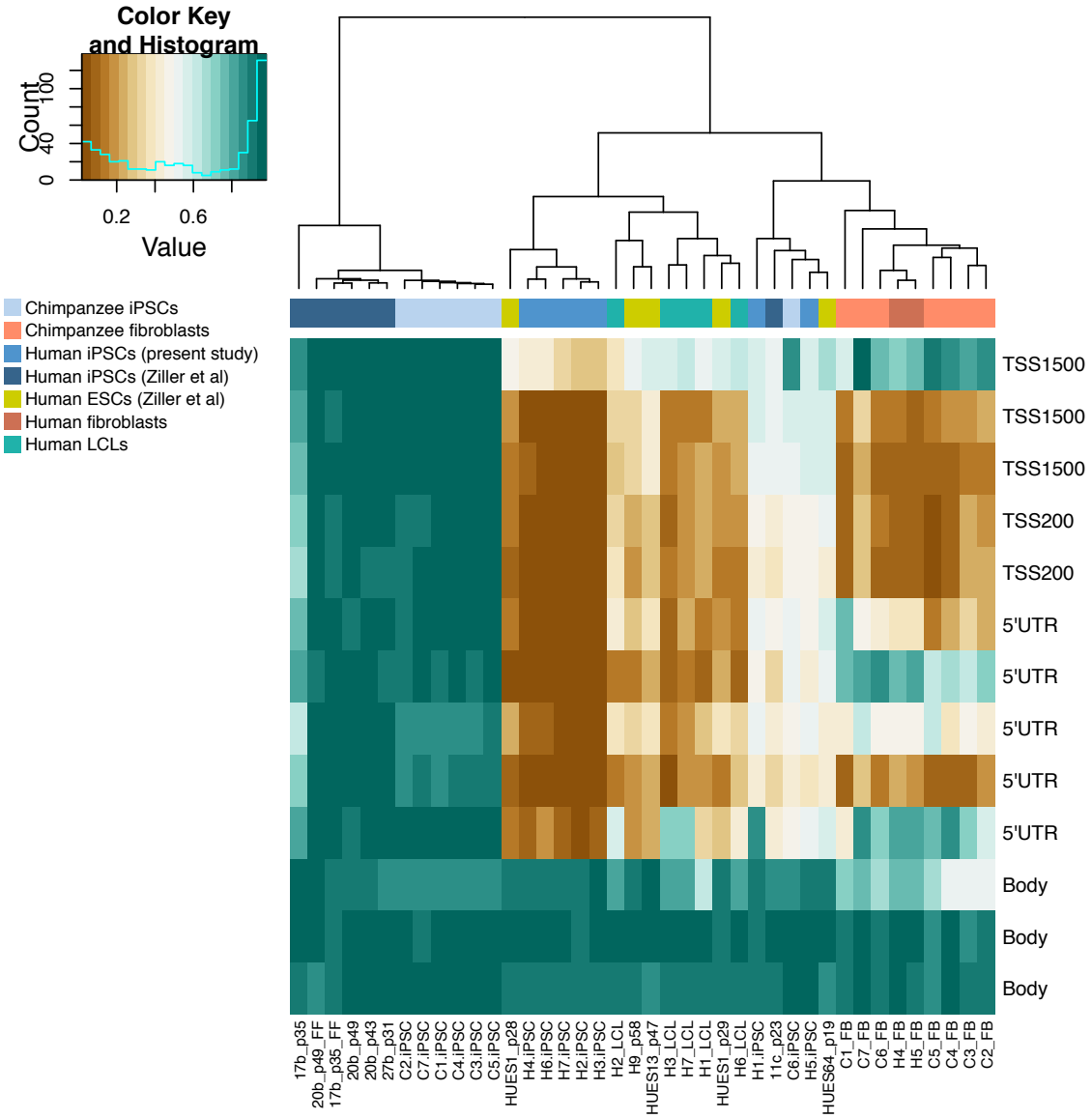


Figure 2.26: **Methylation of REX1 associated CpG sites.** Methylation status of 13 CpG sites associated with *REX1* in chimpanzee and human iPSCs from this study and human PSCs from [175]. *REX1*-expressing chimpanzee iPSC line is colored red; location of the probes relative to the gene sequence is indicated along the right hand side.

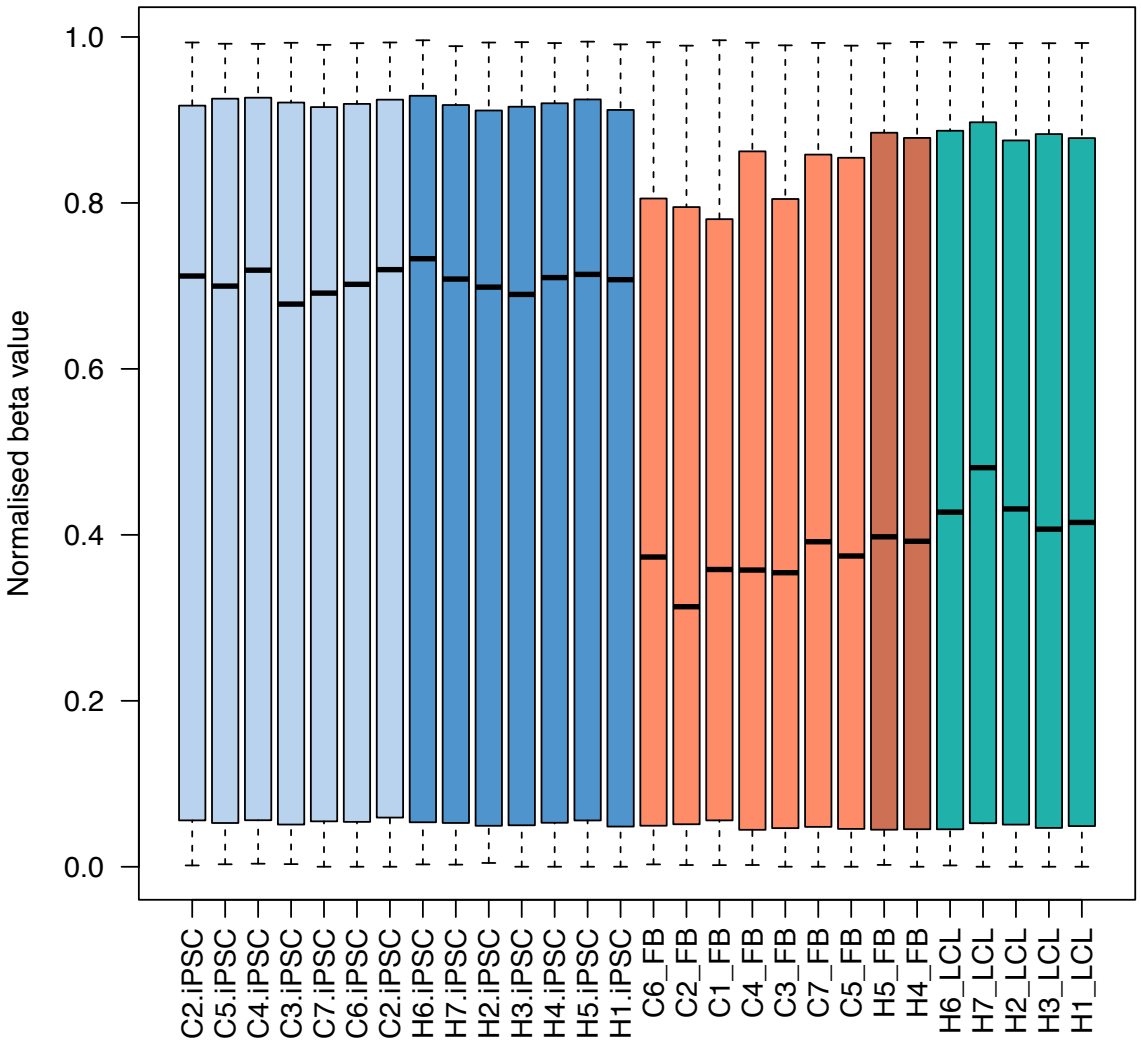


Figure 2.27: **Boxplots of methylation beta values at 335,307 probes across all samples.** Plots are colored by tissue type: light blue: chimpanzee iPSCs; dark blue: human iPSCs; light orange: chimpanzee fibroblasts; dark orange: human fibroblasts; turquoise: human LCLs.

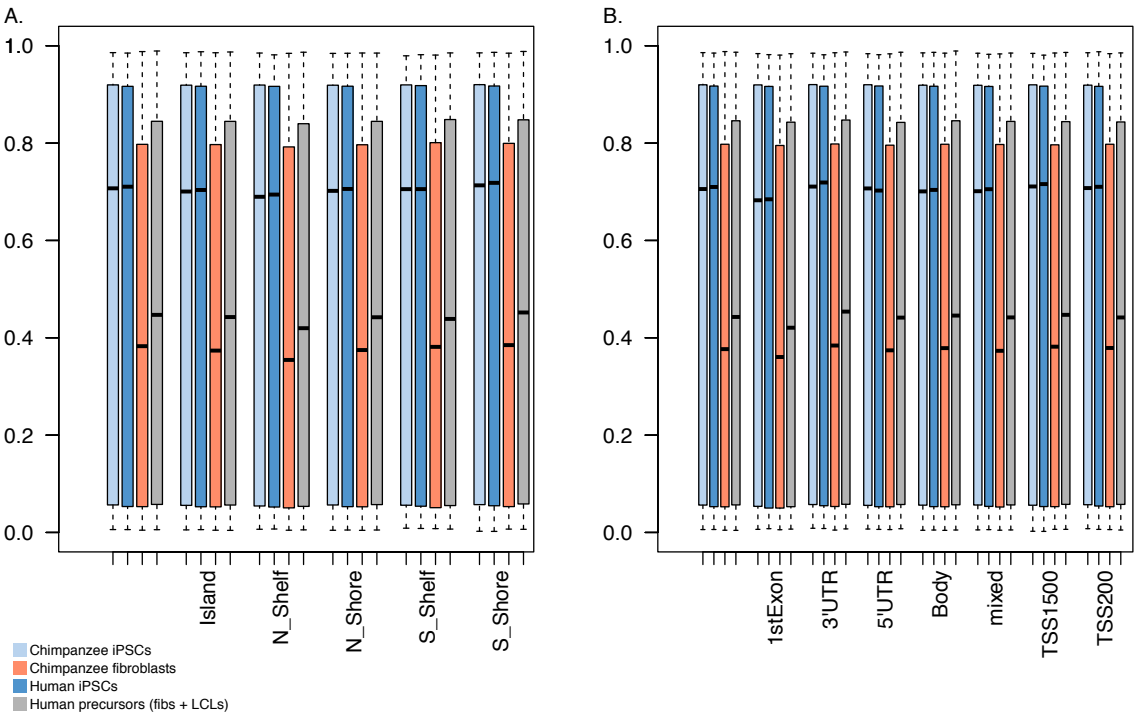


Figure 2.28: **Boxplots of methylation beta values across all samples, grouped by potency and genomic features.** Boxes are colored by tissue type: light blue: chimpanzee iPSCs; light orange: chimpanzee fibroblasts.

Figure 2.29: **Venn diagrams showing overlap in interspecies differences before and after reprogramming.** a. Overlap in DE genes between chimpanzee and human fibroblasts, and chimpanzee and human fibroblast-derived iPSCs. b. Overlap in DM probes between chimpanzee and human fibroblasts, and chimpanzee and human fibroblast-derived iPSCs.

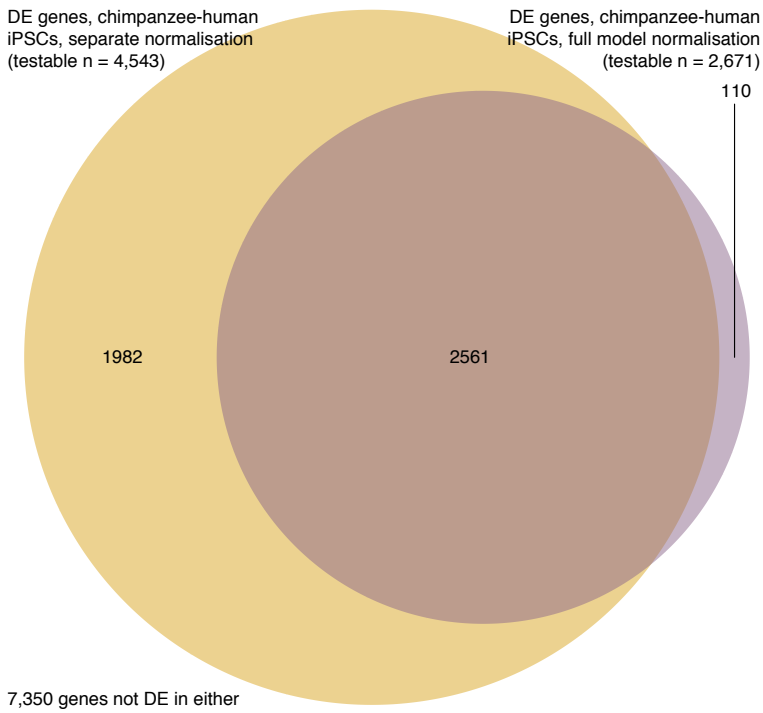


Figure 2.30: **Venn diagram showing overlap of genes identified as DE between iPSCs of the two species when we normalize the iPSC data independently and alongside data from the precursors.**

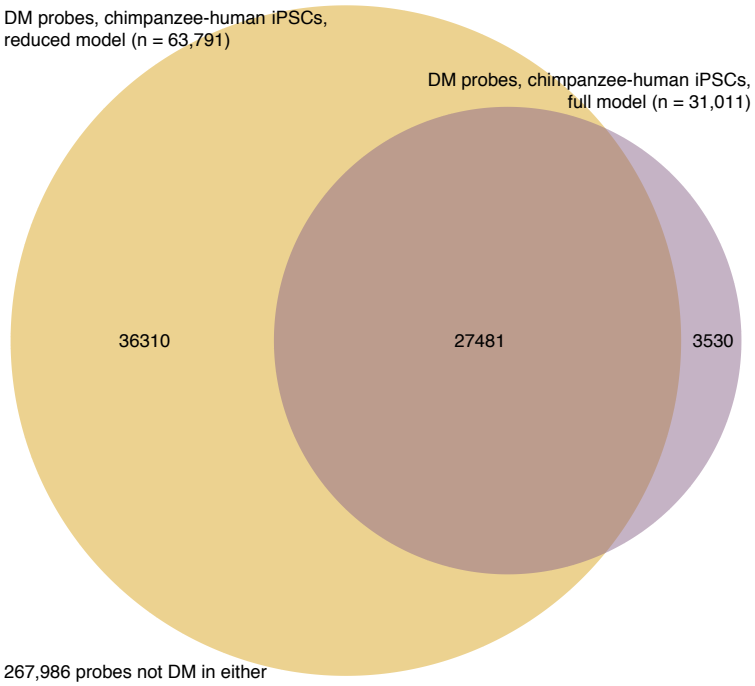


Figure 2.31: **Venn diagram showing overlap of probes identified as DM between iPSCs of the two species under the full and reduced *limma* models.**

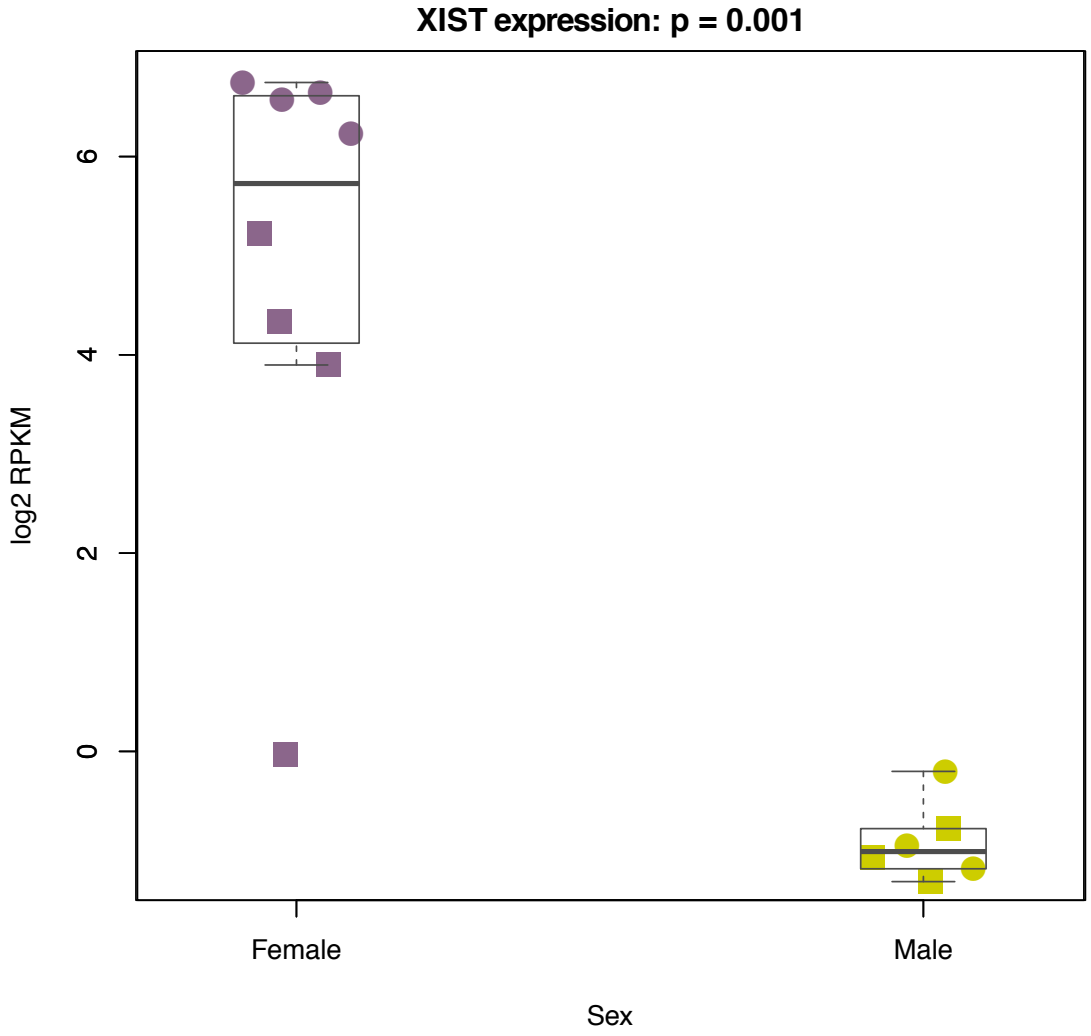


Figure 2.32: **Normalized *XIST* expression values in 7 chimpanzee and human XIST.** Circles denote chimpanzee iPSCs, squares indicate human iPSCS.

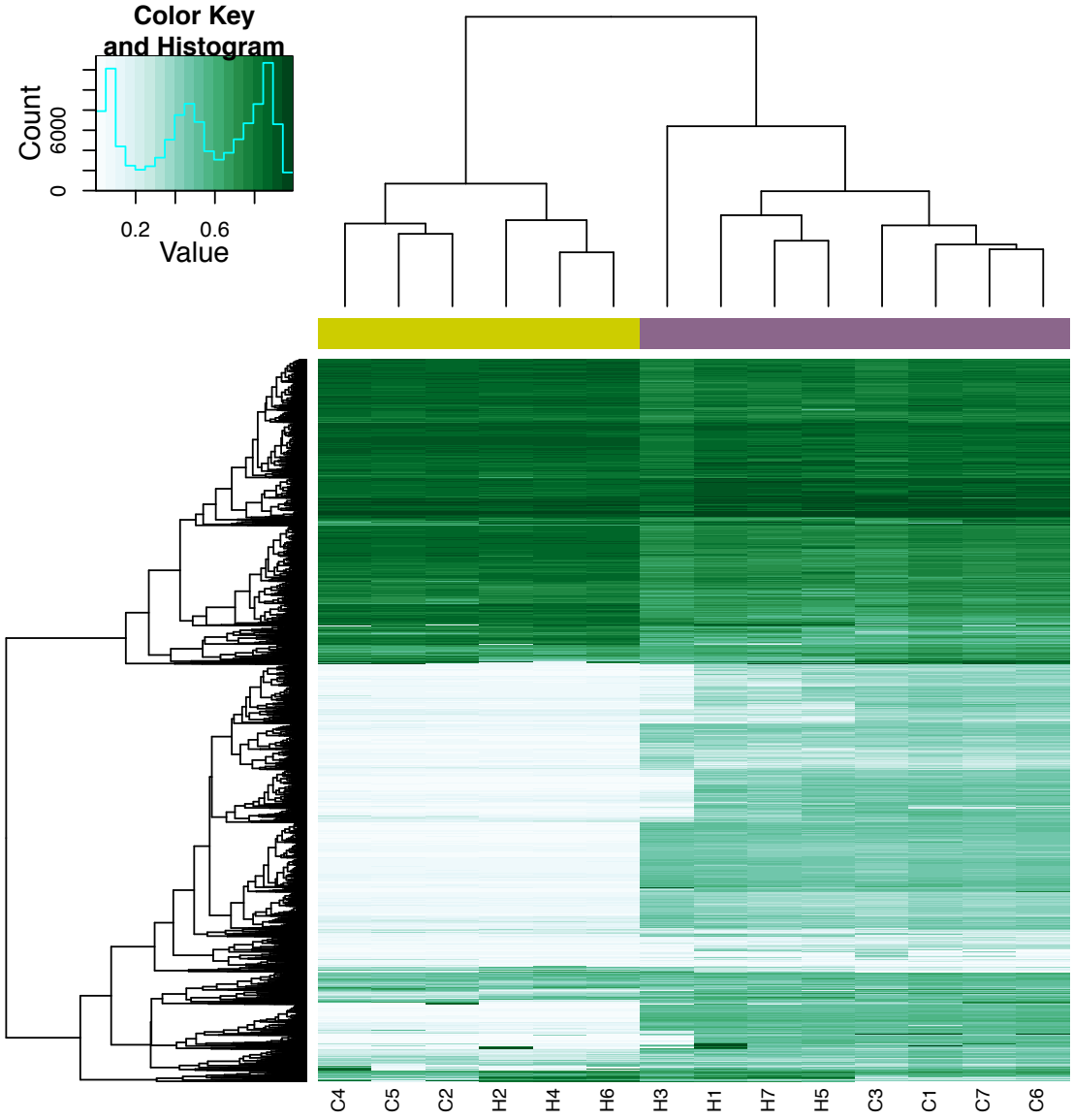


Figure 2.33: **Quantile-normalized methylation beta values at 8,210 X-chromosome probes in 7 chimpanzee iPSCs and 7 human iPSCs.** The color bar beneath the dendrogram indicates sex of the individuals: purple: female; yellow: male. Sample names ending with _ FB indicate fibroblast lines used to generate the corresponding iPSC line, samples ending with _ LCL indicate _ LCL lines used to generate the corresponding iPSC line.

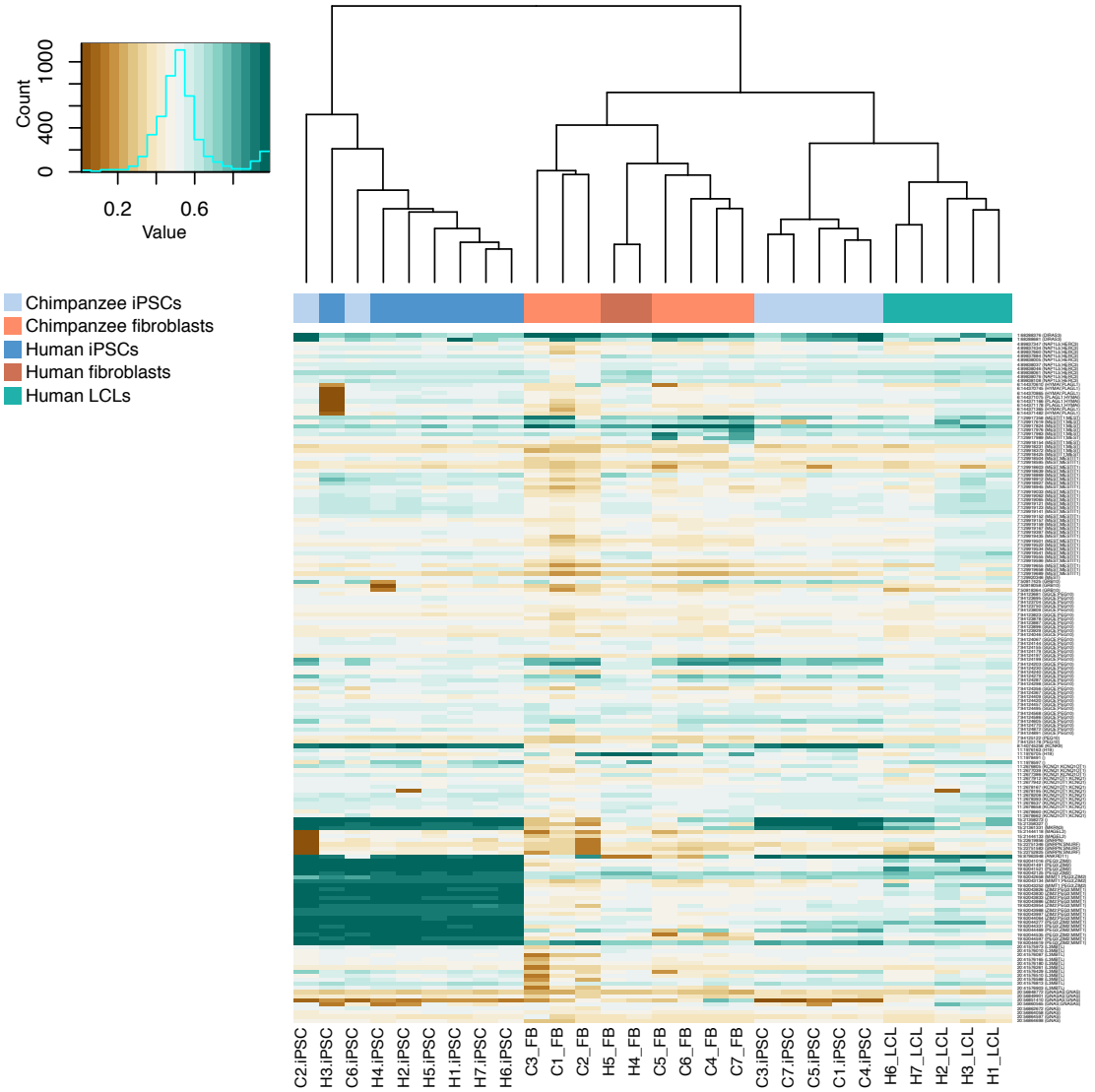


Figure 2.34: **Normalized methylation beta values at 168 assayable probes known to be subject to parental imprinting effects, from [75].** Sample names ending with _ FB indicate fibroblast lines used to generate the corresponding iPSC line, samples ending with _ LCL indicate LCL lines used to generate the corresponding iPSC line.

2.5.2 Supplementary Tables

Table 2.1: **Description of samples.** a: Descriptive data for all chimpanzee cell lines used in this work. b: Descriptive data for all human iPSC lines used.

Table 2.2: **Origin and purpose of all primers used.** For primers not designed in this study PMID is provided

Table 2.3: **Normalized RPKM values and DE genes between chimpanzee and human iPSCs.**

Table 2.4: **Gene Ontology BP terms associated with genes DE between chimpanzee and human iPSCs.**

Table 2.5: **DMRs identified between chimpanzee and human iPSCs.** Coordinates provided for probe start and end based on hg19 genome build

Table 2.6: **Genome-wide data summary statistics.** a: Numbers of DM probes and DMRs between chimpanzee and human iPSCs identified under various mean βdifference thresholds. b: RNA- and ChIP-sequencing reads generated and mapped for all samples in this work. c: Effects of different normalization schemes on the number of genes classified as DE in the full data set.

Table 2.7: **Gene Ontology BP terms associated with genes within DMRs between chimpanzee and human iPSCS.**

Table 2.8: **H3K27ac and H3K27me3 enrichment scores in 3 chimpanzee and 3 human iPSCs around 26,115 orthologous TSSs.**

Table 2.9: **Correlations between principal components and selected covariates.** a: in the expression data. b: in the methylation data.

Table 2.10: **Normalized RPKM values and DE genes identified under the full *limma* DE testing framework.**

Table 2.11: **DMRs identified between chimpanzee iPSCs and their precursor fibroblasts.**

Table 2.12: **DMRs identified between human iPSCs and their precursor cells.**

CHAPTER 3

A COMPARATIVE STUDY OF ENDODERM DIFFERENTIATION IN HUMANS AND CHIMPANZEES

3.1 Abstract

There is substantial interest in the evolutionary forces that shaped the regulatory framework in early human development. Progress in this area has been slow because it is difficult to obtain relevant biological samples. Inducible pluripotent stem cells (iPSCs) may provide the ability to establish *in vitro* models of early human and non-human primate developmental stages. Using matched iPSC panels from humans and chimpanzees, we comparatively characterized gene regulatory changes through a four-day timecourse differentiation of iPSCs (day 0) into primary streak (day 1), endoderm progenitors (day 2), and definitive endoderm (day 3). As might be expected, we found that differentiation stage is the major driver of variation in gene expression levels, followed by species. We identified thousands of differentially expressed genes between humans and chimpanzees in each differentiation stage. Yet, when we considered gene-specific dynamic regulatory trajectories throughout the timecourse, we found that at least 75% of genes, including nearly all known endoderm developmental markers, have similar trajectories in the two species. Interestingly, we observed a marked reduction of both intra- and inter-species variation in gene expression levels in primitive streak samples compared to the iPSCs, with a recovery of regulatory variation in endoderm progenitors. The reduction of variation in gene expression levels at a specific developmental stage, paired with overall high degree of conservation of temporal gene regulation, is consistent with the dynamics of a conserved developmental process.

3.2 Introduction

Differences in gene regulation between humans and other primates likely underlie the molecular basis for many human-specific traits [58]. For example, it has been hypothesized that human-specific gene expression patterns in the brain might underlie functional, developmental, and perhaps cognitive differences between humans and other apes [165, 92]. Providing a measure of support for this notion, a recent comparative study that explored the temporal dynamics of gene regulation found potential differences in the timing of gene expression in the developing brain across primates [130]. The authors argued that such differences might be related to inter-species differences in the timing of developmental processes. Other comparative studies of gene regulatory phenotypes in primates, while challenging, have resulted in important insights into the evolution of gene expression levels and the traits they are associated with [119]. Yet, we are also finding that gene expression patterns alone (without additional context or perturbation) provide little insight into adaptive phenotypes, molecular mechanisms, or even the specific biological processes involved in the observed changes in gene expression levels. To a large degree, the challenge is that comparative studies in humans and non-human apes are extremely restricted because we only have access to a few types of cell lines and to a limited collection of frozen tissues [119]. A few studies have chosen to sidestep this limitation by using model organisms in an attempt to recapitulate inter-primate differences in gene regulation. In this approach, the phenotypic impact of differences in gene regulation between humans and non-human primates are studied with high spatial and temporal resolution in a model species, such as the mouse [40, 104, 105, 126]. These studies are useful and often informative, but when model organisms are used to recapitulate gene regulatory differences between primates, the inference about function requires one to make a critically important assumption, which is typically not tested. Namely, one must assume that the effects of gene regulatory changes on complex phenotypes are identical in model organisms and in humans. This is a common assumption, not exclusively made in

the context of comparative studies, but shared to some extent with every study of human disease that makes use of model organisms. The difference is that model organisms studies of human disease ultimately have to seek evidence that inference based on model organisms is relevant to humans. In contrast, studies of human evolution using model organisms rarely, if at all, are required to meet this standard. Indeed, oftentimes it is unclear how to design an experiment in model organisms that can directly address a phenotype difference between humans and non-human primates, for instance, when the phenotypes under consideration are related to cognitive abilities. In such cases, the assumption that the effects of gene regulatory changes on complex phenotypes are identical in model organisms and in humans is somewhat strained [40]. The caveats associated with using a model organism to study the phenotypic effects of regulatory differences between primates notwithstanding, until recently, it is not clear that there was an alternative approach. While current comparative studies using primate material (tissue samples or cell lines) have provided valuable insight into the genetic architecture of gene regulation, we did not have a flexible and faithful framework with which to dynamically study inter-species variation in gene regulation [119]. In particular, frozen post-mortem tissues are not optimal templates for many functional genomic assays; as a result, we lack datasets that survey multiple dimensions of gene regulatory mechanisms and phenotypes from the same individuals [13, 119]. Moreover, because it is rare to collect a large number of tissue samples from the same donor (particularly in non-human primates), we have never had the opportunity to study cross-species, population-level patterns of gene regulation in multiple tissues or cell types derived from the same genotype (same donor) in non human apes. We also have not been able to comparatively study population-level dynamics of gene regulation in primates, for example, during perturbation. In order to gain true insight into regulatory processes that underlie variation in complex phenotypes, we must have access to faithful model systems for a wide range of tissues and cell types. In other words, to utilize comparative functional approaches to comparatively study the

genetic architecture of complex phenotypes in humans and other apes, a new approach is needed. Recent technological developments in the generation and differentiation of induced pluripotent stem cells (iPSCs) now provide a renewable, staged and experimentally pliable source of terminally differentiated cells. Utilizing timecourse differentiation protocols, we can examine the context dependent nature of gene regulation, as well as the temporal roles of gene expression as different cell types and developmental states are established [73]. This approach seems promising, and indeed, a handful of recent studies have been successful in utilizing iPSCs from humans and chimpanzees to characterize the uniquely human aspects of craniofacial development [106] and cortex development [86, 94]. Primate iPSC panels are a particularly attractive system for comparative studies of early development. Based on studies in model organisms, we expect differentiation into a germ layer in a mammalian system to be extremely conserved [176]. We thus set out to ask whether we can recapitulate a conserved early developmental gene regulatory trajectory in human and chimpanzee iPSC-derived cell lineages. Our rationale is that if we can demonstrate the fidelity of the iPSC model in this context, this would provide support for the notion that the iPSC system can be a useful tool for future comparative studies of dynamic biological processes. To this end, we chose to differentiate iPSCs from human and chimpanzee into the endoderm germ layer; from which essential structures in the respiratory and digestive tracts are ultimately derived, 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 [176]. Using this system, we found evidence to support developmental canalization of gene regulation in both species, 24 hours after differentiation from an iPSC state.

3.3 Results

3.3.1 Study design and data collection in the iPSC-based system

To perform a comparative study of differentiated cells, we used a panel of 6 human and 4 chimpanzee iPSC lines previously derived and characterized by our lab [30, 42]. We differentiated the iPSCs into definitive endoderm, a process that was completed over 3 days [73], and included replicates of cell lines that were independently differentiated (see Methods; Fig. 3.1A). We performed the differentiations in two batches, with equal numbers of human and chimpanzee samples in each batch. We harvested RNA from iPSCs (day 0) prior to differentiation and subsequently every 24 hours to capture intermediate cell populations corresponding to primitive streak (day 1), endoderm progenitors (day 2), and definitive endoderm (day 3). Overall, we collected a total of 32 human samples and 32 chimpanzee samples (Fig. 3.1A). We confirmed that RNA from all samples was of high quality (3.1; Fig. 3.10) and subjected the RNA to sequencing to estimate gene expression levels. Detailed descriptions of all individual donors, iPSC lines, sample processing and quality, and sequencing yield, can be found in the Methods section and 3.2. To estimate gene expression levels, we mapped reads to the corresponding genome (hg19 for humans and panTro3 for chimpanzees) and discarded reads that did not map uniquely [57]. We then mapped the reads to a list of previously described metaexons across 30,030 Ensembl genes with one-to-one orthology between human and chimpanzee [13, 15]. We eliminated genes that were lowly expressed in either species, removed data from one clear outlier sample (H1B at Day 0; Fig. 3.11A), and normalized the read counts (see Methods; Fig. 3.11B) to obtain TMM- and cyclic loess-normalized $\log_2$ counts per million (CPM) values for 10,304 orthologous genes (Fig. 3.2; 3.3). These normalized gene expression values were used in all downstream analyses.

Beyond the gene expression data we collected, in the second batch, we assessed the purity of the cell cultures at each day of the timecourse using flow cytometry with a panel of six

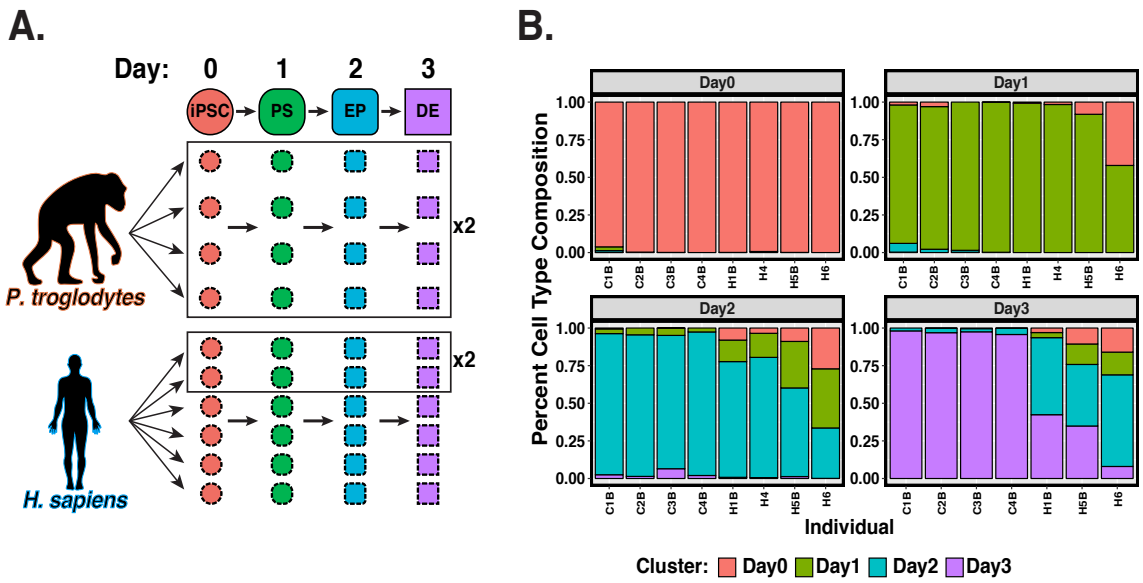


Figure 3.1: **Study design and purity.** A. Study design. Samples from four chimpanzees and six humans were studied at four time-points during endoderm development. We included two technical replicates from each of the chimpanzees and two technical replicates for two of the six humans. iPSC: induced pluripotent stem cell, PS: primitive streak, EP: endoderm progenitor, DE: definitive endoderm. B. Purity analysis: Cell type composition at each day based on FACS analysis (see methods), estimated by k-means clustering.

canonical markers. These markers correspond to the cell types we expected in the different stages of differentiation (Fig. 3.1B; Fig. 3.12;3.4). We assessed purity of the samples in the first batch of differentiation as well, but due to a technical problem with the antibodies, those values are not informative. Overall, the FACS-based estimated purity levels for the sample in the second batch are high and consistent across species in days 0 and 1 (> 0.79 and > 0.66, respectively; Fig. 3.1B;3.4; Fig. 3.12-Fig. 3.13). In days 2 and 3, however, purity levels were considerably lower in the human than in the chimpanzee samples (Fig. 3.1B;3.4; Fig. 3.12-Fig. 3.13). On the one hand, this inter-species difference in purity in the later stages of the time course limits the insight we can draw from this comparative study, as we discuss throughout the paper. On the other hand, because our main focus is to confirm that this early developmental process is conserved in the two species, the observation of strong conservation, which we discuss below, is robust with respect to this inter-species technical difference in purity.

3.3.2 iPSCs-based system effectively models primate endoderm differentiation

Given the potential impact of study design properties on gene expression data and subsequent conclusions [166], as a first step of our analysis, we confirmed that none of our recorded variables related to sample processing (apart from sample purity, as stated above) were confounded with our main variables of interest, namely day and species (see Methods;3.2 and 3.5; Fig. 3.14). Once we were confident that our study design provided an effective data set for addressing our biological questions of interest, we performed a global survey of the gene expression data using principal component analysis. This analysis indicated that the primary sources of gene expression variation are differentiation day (Fig. 3.2A; 3.5; regression of PC1 by differentiation day, $P< 10^{-15}$), followed by species (regression of PC2 by species, $P< 10^{-15}$). This observation was also supported by clustering analysis based on

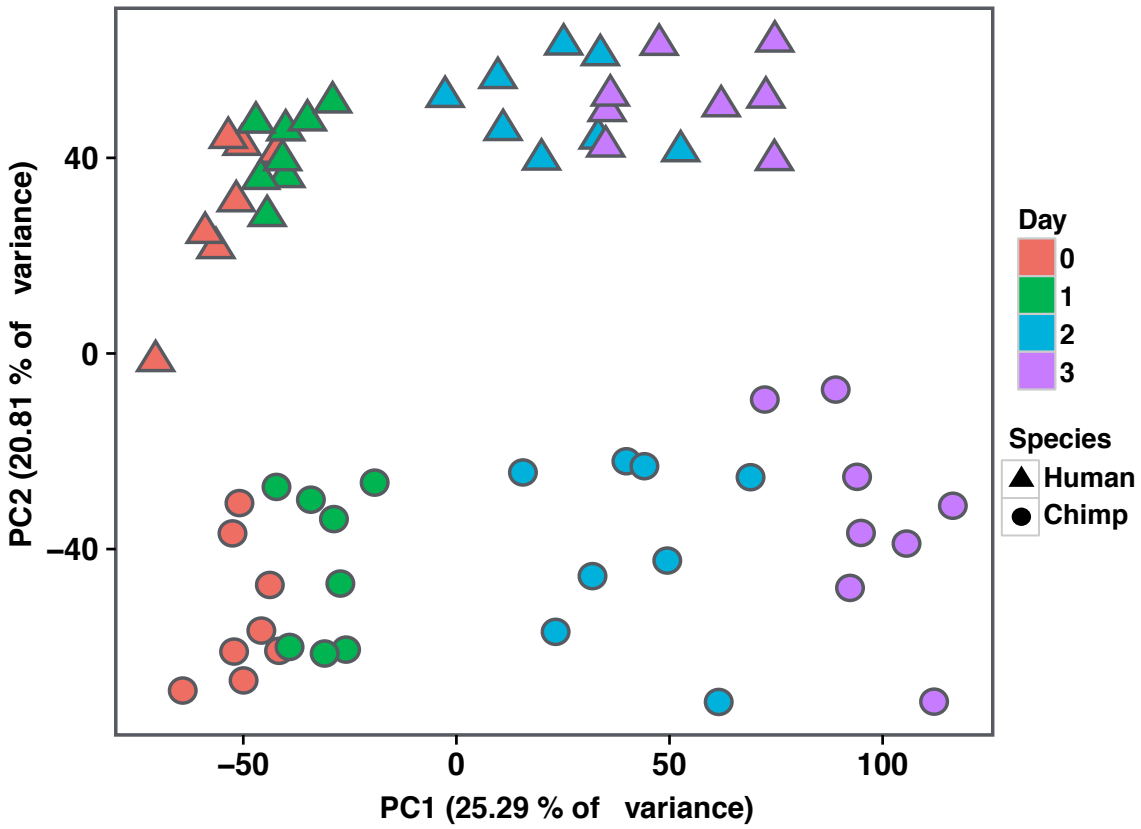


Figure 3.2: **General patterns in the data.** Normalized $\log_2$(CPM) expression measurements for all genes projected onto the axes of the first two principal components. Color indicates day. Shape represents species. PC1 is highly correlated with differentiation day (r = 0.92). PC2 is highly correlated with species (r = 0.93).

the correlation matrix of pairwise comparisons of the gene expression levels (Fig. 3.15).

After characterizing global gene expression patterns, we focused on the expression of specific transcription factors with known roles in developmental pathways (Fig. 3.3A) and other previously known lineage specific markers [35, 73, 139]. Consistent with the results of our FACS analysis, we observed that the temporal trajectory of expression levels of known lineage specific markers and transcription factors further supported the assumed differentiation stages in each day (e.g. primitive steak-specific markers had increased expression on day 1, Fig. 3.2B). The lineage specific markers and transcription factors were expressed at comparable levels in humans and chimpanzees at the relevant time points, consistent with

previous literature [73, 139], and further supporting the validity of our *in vitro* system (Fig. 3.16A).

3.3.3 Comparative assessment of gene expression changes during differentiation

To identify gene expression differences between humans and chimpanzees throughout the timecourse, we used the framework of linear models (see Methods). We first assessed how many genes were differentially expressed (DE) between species at each time point independently. Using this approach, we classified thousands of genes as DE between the species (at FDR of 5% 4475 5077 genes are classified as inter-species DE at different times points; Fig. 3.4A; 3.6). Even at a fixed FDR cutoff, nearly half of the genes that were classified as DE between the species at any single time point were found to be DE in all time points (2269 genes). Nearly a third of genes whose expression was measured in our experiment were not classified as DE between the species at any time point (2862 genes, 28%).

We proceeded to consider temporal expression patterns within species. When analyzing expression changes across consecutive time points, we had more power to detect temporal gene expression differences in chimpanzee compared to humans (Fig. 3.4B-C and Fig. 3.5; 3.7), especially with respect to the transition between endoderm progenitors (day 2), and definitive endoderm samples (day 3). This property is likely related to the inter-species difference in purity of samples from these days, as discussed earlier. When we accounted for incomplete power (see Methods), we found a remarkably consistent pattern whereby 77% of DE genes between iPSCs and primitive streak in humans are also DE between these states in chimpanzees; similarly, 77% of DE genes between primitive streak and endoderm progenitors in humans are also DE between these states in chimpanzees; and 80% of DE genes between endoderm progenitors and definitive endoderm in humans are also DE between these states in chimpanzees (3.8). As might be expected from these observations, we found that

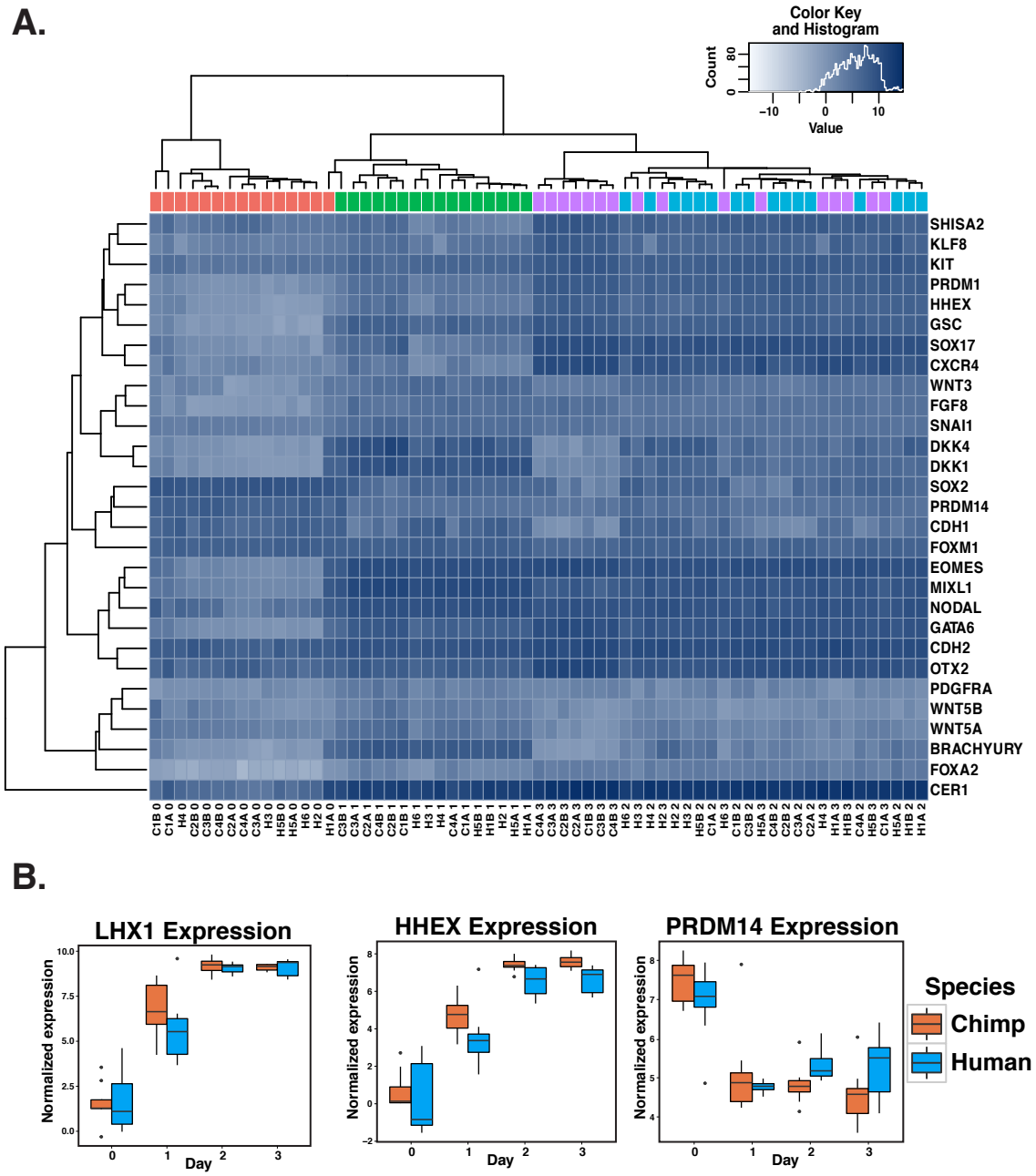


Figure 3.3: **Canonical markers of differentiation.** A. Heat map of normalized $\log_2$ (CPM) as a measure of expression levels of transcription factors that are known to be highly expressed in one or more stages in the differentiation to endoderm [73]. Generally, samples from the same day, regardless of species, cluster together. B. Box plots of normalized expression values for genes with known roles in endoderm development.

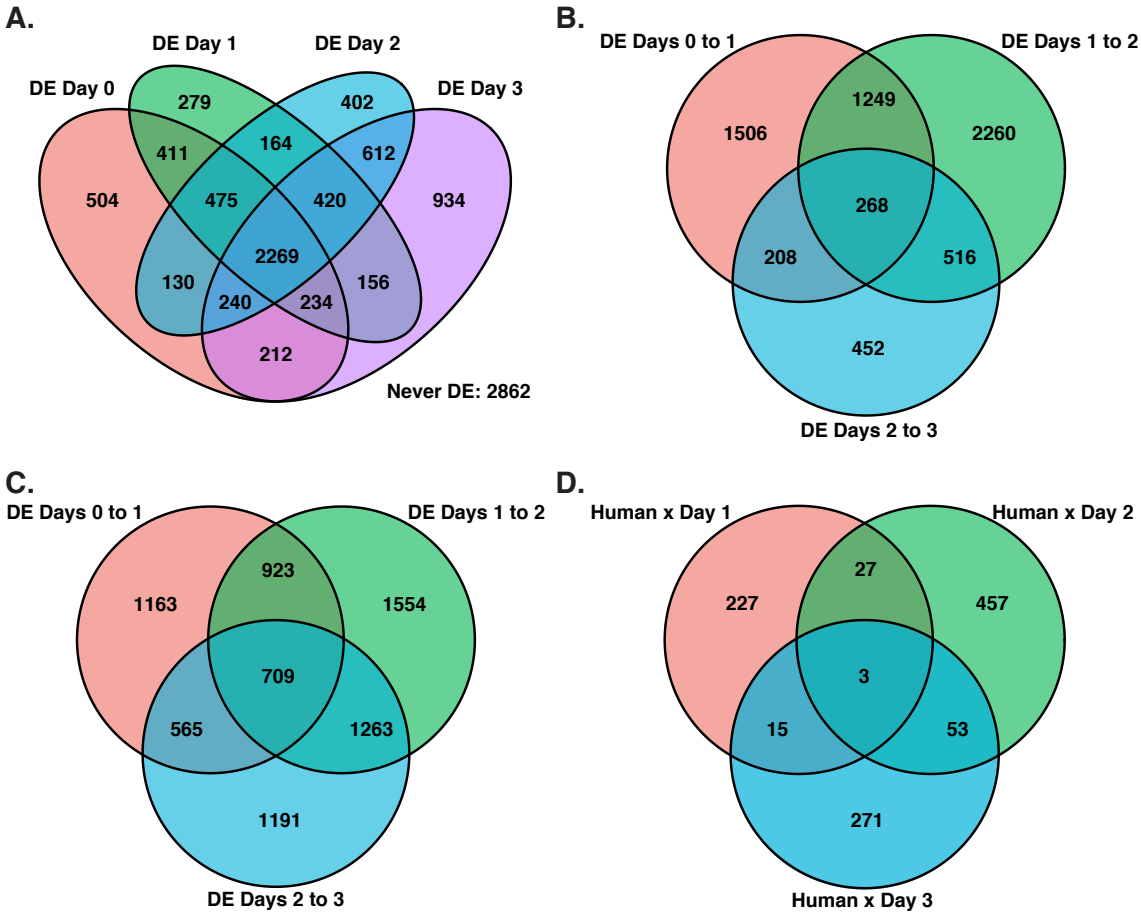


Figure 3.4: **Number of differentially expressed (DE) genes in pairwise analyses.** Venn diagrams of A. DE genes at each day B. DE genes between consecutive time points in humans C. DE genes between consecutive time points in chimpanzees D. genes with a significant species-time point interaction effect at each day (DE was classified at FDR of 5% in all cases).

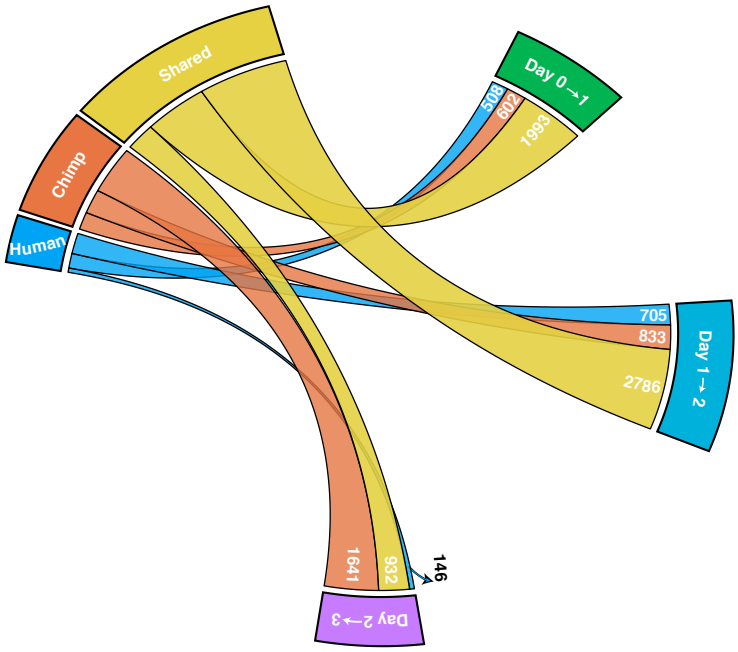


Figure 3.5: **High sharing of DE genes across species.** A circos diagram with the number of shared, human-specific, and chimpanzee-specific DE genes across time points. There is a high degree of sharing of DE genes (yellow ribbon), particularly from day 0 to 1 and day 1 to 2.

the relationship between day and gene expression was largely independent of species (Fig. 3.4D;3.9).

3.3.4 Joint Bayesian analysis reveals conservation of temporal gene expression profile

In an attempt to further overcome issues of incomplete power affecting these original naïve pairwise DE comparisons, and to account for dependency in data from different time points, we utilized a Bayesian clustering approach implemented by Cormotif [155]. This joint modeling technique leverages expression information shared across time points to identify the most common temporal expression patterns (referred to as correlation motifs). We identified diverse expression patterns that emerge as differentiation progresses in both species (Fig. 3.6;3.10) as well as a set of 3789 genes whose expression is not significantly altered

throughout the timecourse (8004 genes could be reliably classified into a motif, 3.10; Fig. 3.17A). This analysis revealed further evidence for conserved gene expression patterns, as 75% of genes assigned were assigned to motifs with the same or similar temporal regulatory trajectories in both species (Fig. 3.6).

Further, when we excluded data from the definitive endoderm samples, where we suspect that a particularly large inter-species difference in sample purity has increased gene expression variance between the species, we assigned 85% of genes to motifs with the same temporal trajectories across species. These observations are robust with respect to the number of correlation motifs (Fig. 3.17B), the method used to combine data from technical replicates (Fig. 3.17C and Fig. 3.18), which days were included in the pairwise comparisons (Fig. 3.17D), and the inclusion of all 10,304 genes in the analysis. Our observations are also robust with respect to the overall approach used to estimate and compare gene expression trajectories (Fig. 3.16B). We found two correlation motifs with a potential marked difference between the species at a given stage (motif 4 with 187 genes and motif 7 with 686 genes). In both of these motifs, data from the earliest time points were conserved but gene regulation in the final stage (day 2 to 3) differed between the species. The genes in these motifs were enriched for Gene Ontology (GO) annotations related to animal organ development (e.g. *NRTN, PITX2, RDH10*), anatomical structure morphogenesis (*ARHGDIA, EHD2, SERPINE1*), regulation of developmental process (*FLRT3, LOXL2, SEMA7A*), and regulation of cell differentiation (*DIXDC1, ENC1, IRF1*;3.10B) [2]. These four GO annotations were not enriched in other similarly sized motifs or group of motifs (3.10C-E) [2]. Unfortunately, due to interspecies differences in purity at the later days, we cannot definitively determine whether these enrichments are driven by biological or technical differences.

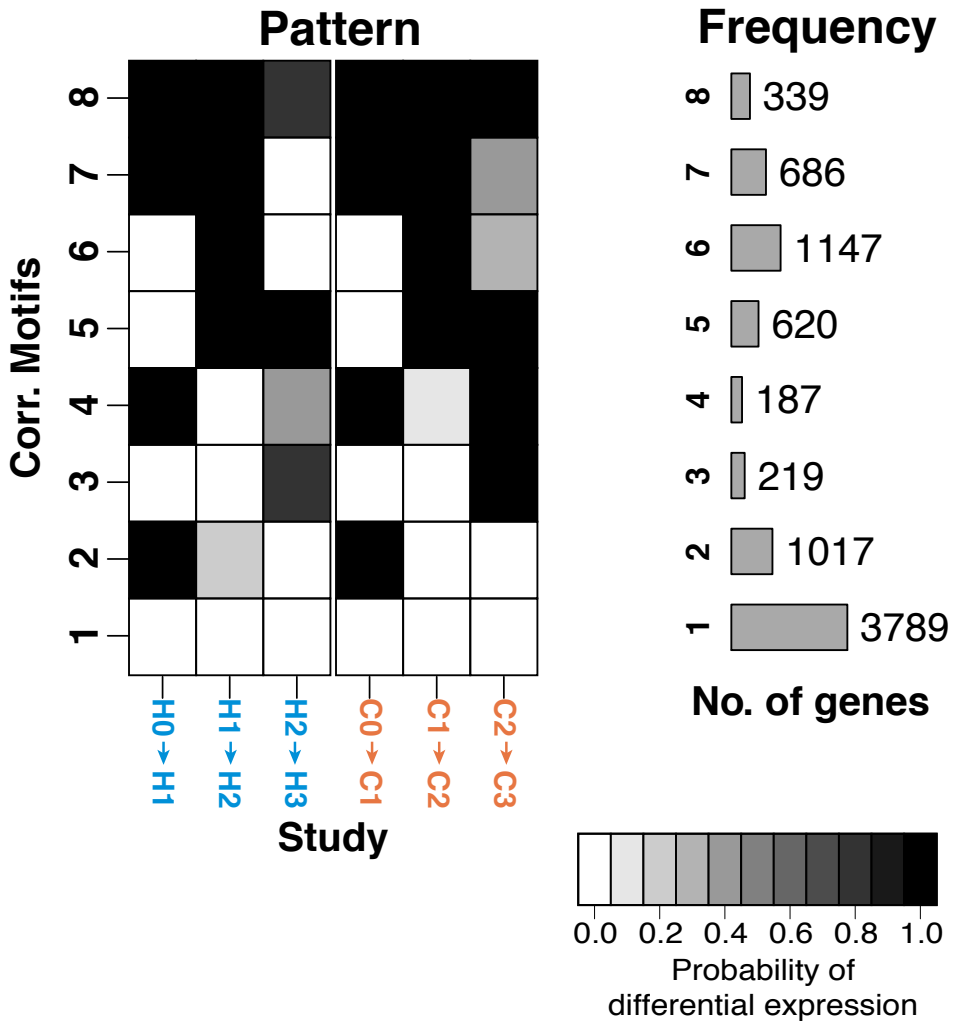


Figure 3.6: **Gene expression motifs.** Correlation motifs based on the probability of differential expression across days for each species with the number of genes assgined to each correlation motif. The shading of each box represents the posterior probability that a gene is DE between two time points in a given species. Each row ("correlation motif") represents the most prevalent expression patterns. 8004 out of 10,304 total genes were assigned to one correlation motif in this model.

3.3.5 Reduced variation in gene expression levels at primitive streak

We next turned our attention to differences in the magnitude of variation in gene expression levels across time points, within and between species. Previous studies reported that variation in gene expression levels between individuals was lower in iPSCs than in differentiated cells ([5] and Fig. 3.19A). We were thus interested in gene expression variation during iPSC differentiation in our comparative system. We first compared within-species expression variation for all 10,304 orthologous genes across time points. We found a reduction in inter-individual variation of gene expression levels as the human samples differentiated from iPSCs to primitive streak ($P< 10^{-15}$, Fig. 3.7). We also detected this pattern when we considered the chimpanzee samples ($P< 10^{-15}$, Fig. 3.7), but the effect size in chimpanzee is much smaller. We did not identify similar reduction in variation in gene expression levels in any other transition during the timecourse in either species ($P> 0.5$ for testing the null of no change in variance of gene expression from day 1 to 2 and from day 2 to 3 in each species). As mentioned above, the purity of the samples in days 2 and 3 is lower than that of samples in days 0 and 1, though we only successfully measured the purity values for the samples that were processed in the second batch. We accounted for the measured purity values by regressing them out of the gene expression data. As might be expected, accounting for the purity values resulted in different patterns observed for the data from days 2-3. Yet, the reduction in variation from day 0 to 1 persisted in both species even after we accounted to the purity values (Fig. 3.19B).

We turned our attention to regulatory divergence between species. The overall human-chimpanzee divergence in gene expression levels was also slightly reduced as samples differentiated from iPSCs to primitive streak (Mann-Whitney U Test, $P= 0.04$; Fig. 3.19C), but not in any other transition during the timecourse. Furthermore, while we classified 504 genes as DE between humans and chimpanzees exclusively in iPSCs (of a total of 4,475 DE genes in iPSCs, FDR = 5%; Fig. 3.4A;3.11), we found only 279 genes that were DE exclusively

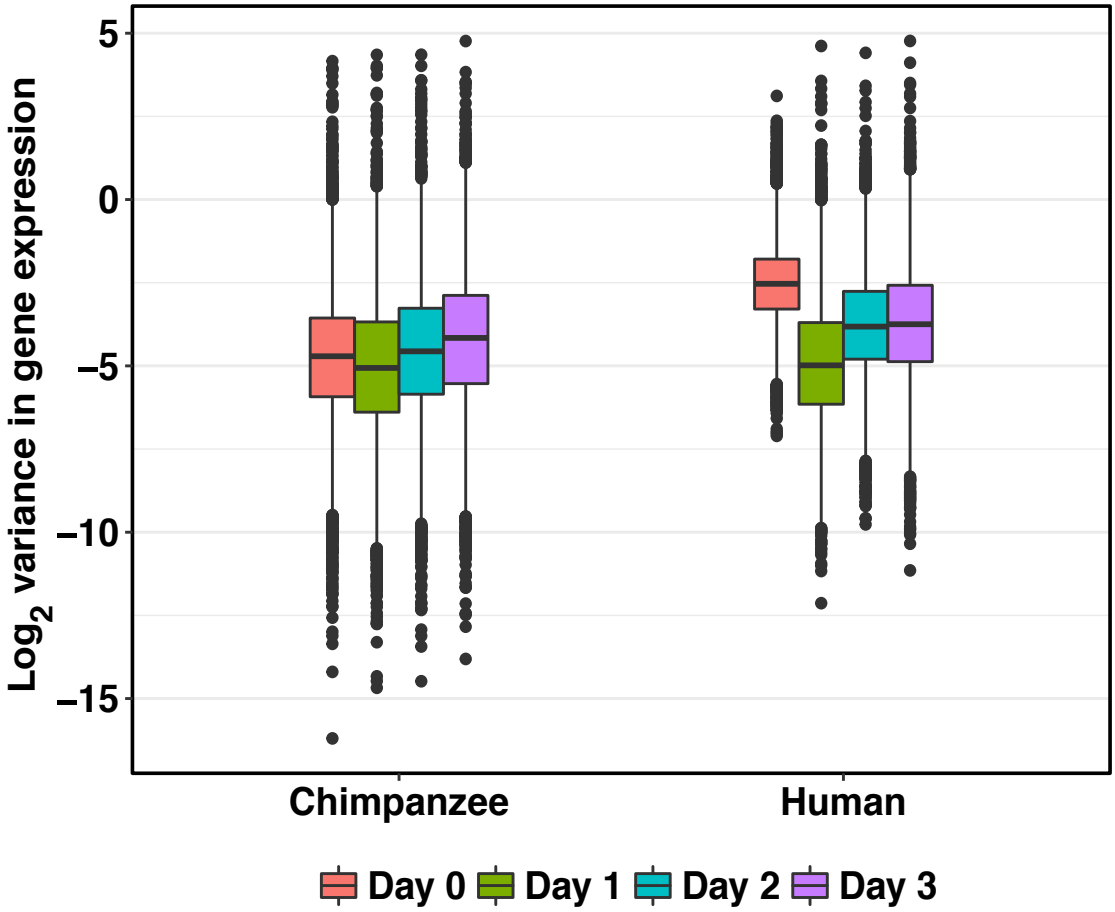


Figure 3.7: **Global reduction of variation in gene expression from the iPSCs to primitive streak state.** Box plot of the $\log_2$ variance of expression levels for each gene. Variation in gene expression levels are significantly reduced from iPSCs to primitive streak (P <10^{-15} in both species) but not in subsequent time points (P >0.5 in both species).

in primitive streak samples (from a total of 4,408 DE genes for the primitive streak; Fig. 3.4A;3.11). The number of genes that are DE between the species exclusively in endoderm progenitors and definitive endoderm samples is higher (at FDR of 5%, 402 and 934, respectively; Fig. 3.4A;3.11). The observation of a smaller number of genes that are DE exclusively in primitive streak samples compared with iPSCs is robust with respect to the FDR cutoff, differentiation batch, and purity of the samples (Fig. 3.11C and Fig. 3.20;3.11). While the difference in divergence and the number of DE genes between these differentiated states is modest, and could potentially be explained by a number of non-biological factors (including the differences in purity in the later day), this observation was intriguing to us. We thus focused on the transition between iPSCs to primitive streak in both species. The recorded technical factors (including purity for these states) are highly similar across biological conditions in days 0 and 1, and therefore are not likely to explain this observation (3.5 and 3.12). We thus proceeded to analyze the trajectory of variation in expression level on an individual gene basis. In this analysis, we were particularly interested to address whether the individual genes that undergo a change in variation of expression levels are shared across species. An observation of excess sharing could not be explained by inter-species differences in technical factors and hence would provide substantial support to the notion of conserved reduction of regulatory variation as the samples begin their differentiation process. We used F tests to identify genes whose within-species variation in expression levels differs across time points (see Methods). Distributions of Pvalues from all tests can be found in Fig. 3.8A-B and 3.13, which indicate that for a large number of genes, within-species variation in expression levels were reduced exclusively in primitive streak samples. Indeed, while we did not have much power to detect differences in variation of individual gene expression levels between states (due to the small number of individuals in each species), we observed a clear excess of small Pvalues when we tested the null hypothesis that there was no reduction in gene expression levels from day 0 to 1. Using Storeys approach [135] to account for incomplete power, we

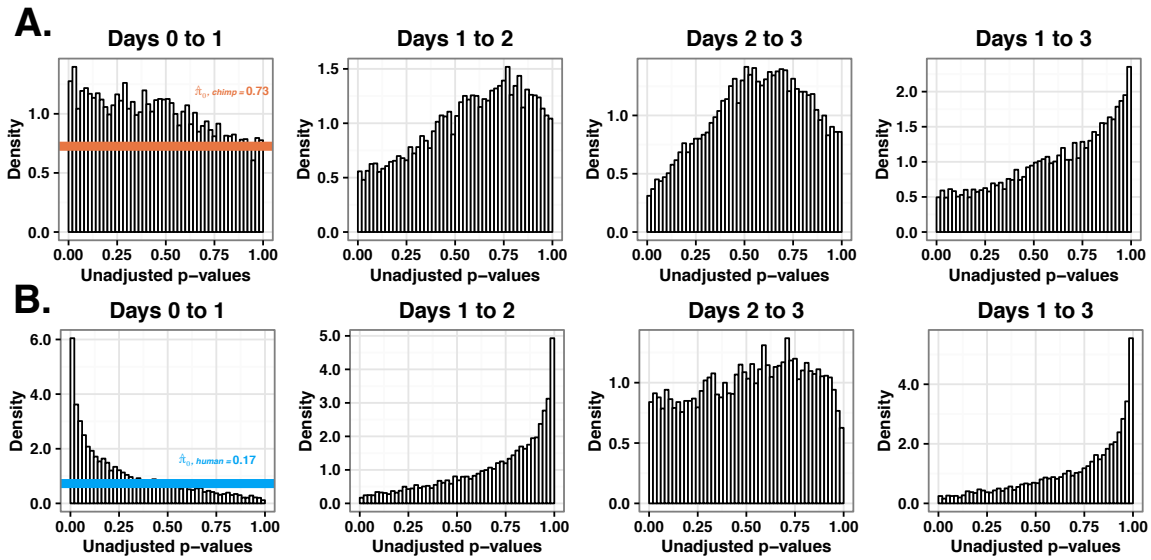


Figure 3.8: **Reduction of variance observed at transition from day 0 - day 1.** A. P value distributions of F tests under the null hypothesis that there is no reduction in variation in gene expression levels as samples progress along the time course in human samples. B.Same as in A, but for Chimpanzee samples. $\hat{\pi}_0$ is the estimated proportion of null tests in each distribution.

estimated that within-species variation in expression levels was reduced as the samples differentiate from iPSCs to primitive streak in 83% and 27% of human and chimpanzee genes, respectively (Fig. 3.8A-B). This result was robust with respect to the method used to calculate the proportion of true positives [136] (Fig. 3.21). We did not observed reduced variation of gene expression in any other differentiation state in our data (Fig. 3.8;Fig. 3.22A).

We next asked about the overlap of genes with reduced variation in primitive streak samples across the two species. Specifically, we asked whether human genes with lower within-species variation in expression levels in primitive streak are more likely to show the same pattern in chimpanzee genes. For this analysis, we again used the Storey approach [135] to estimate the proportion of true positive tests in one species, conditional on the observation of reduced variation in the other species (see Methods). We estimated that 47% of genes whose variation in expression level is reduced in human primitive streak samples showed a similar pattern in chimpanzees (under a permuted null we expect 27%, $P < 10^{-4}$,

Fig. 3.9C;3.14). When we condition on observing a reduction of variation in chimpanzees, the overlap with humans was 84% (under a permuted null we expect 83%, P= 0.38; Fig. 3.9D). This high value was not unexpected because of the initial large proportion of human genes with a clear signature of reduced variation in primitive streak. Because any technical differences that are confounded with species would contribute to increased inter-species differences, these observations support, yet probably underestimate, the degree of high conservation of regulatory patterns in humans and chimpanzees. Using a similar approach to account for incomplete power, we also found a marked overlap of genes whose expression underwent a significant increase in variation throughout the transition from primitive streak to endoderm progenitors (Fig. 3.97E-F; Fig.3.22B). All our observations were robust to a wide range of statistical cutoffs used to classify genes whose within-species variation changes across the differentiation states (Fig. 3.23-Fig. 3.25). In particular, because these are reports of conserved patterns, they are likely to underestimate the degree of conservation given technical differences between the species, including the difference in purity of the samples in days 2 and 3.

Finally, we sought to provide insight into the potential functional consequences of our observations. Interestingly, genes that show reduction of variation in gene expression levels in both species are enriched for GO annotations [2] related to development, including neuroepithelial cell differentiation (e.g. *MYCL, NODAL, CDH2*), cell migration during gastrulation (*FGF8, MIXL1*), and trophectodermal cell differentiation (*CNOT3*, *EOMES* , *SP3*;3.13C). Moreover, we found that null mutation in mouse orthologs of the primate genes with shared reduction of expression variation in our study were more highly associated with embryonic lethality than null mutation in mouse orthologs of primate genes that did not show that pattern (41% versus 28%; $P< 10^{-4}$;3.13D-E).

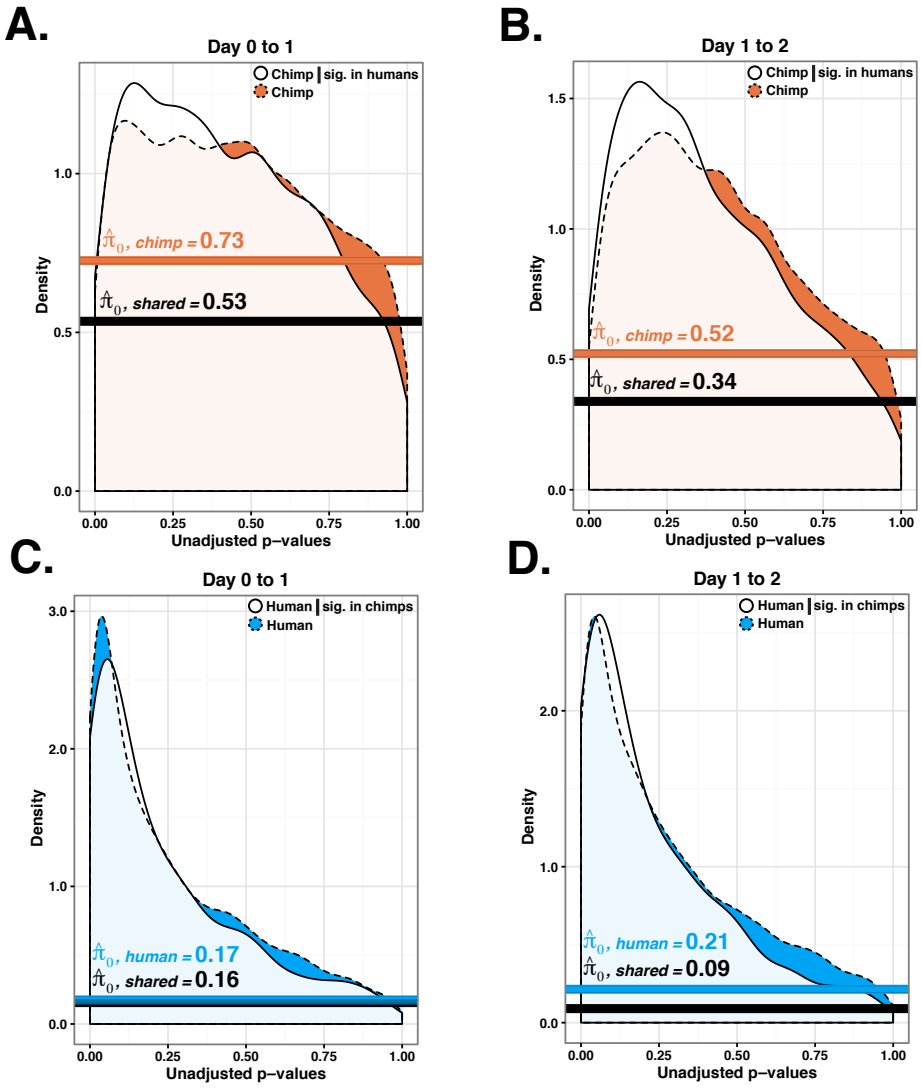


Figure 3.9: **Conserved patterns of reduced variation in gene expression at primitive streak.** A. P value distributions of F tests (from Fig. 3.8 for day 0 to 1 in chimpanzees only for genes who variation was classified as reduced (P <0.05) in humans. B. same as A, but for day 1 to 2. C. P value distributions of F tests (from Fig. 3.8 for day 0 to 1 in humans only for genes who variation was classified as reduced (P <0.05) in chimpanzee. D. same as C, but for day 1 to 2.

3.3.6 Discussion

Our results indicate a strongly conserved temporal expression profile across species during early differentiation. We observed a large number of genes with similar expression profiles across species. Indeed, we found that DE genes between differentiation states are shared between the two species far beyond what expected by chance alone. When we jointly analyzed data from the entire timecourse, nearly all the gene expression trajectory motifs we identified, including 75% of all genes assigned to a motif, are shared across the two species (Fig. 3.6). Still, our observations likely underestimated the proportion of shared regulatory patterns due to incomplete power. Our finding that regulatory trajectories throughout endoderm differentiation are generally highly conserved in these two species was expected. Yet, our observation that a large number of genes are associated with conserved reduced regulatory variation in a specific transition state is a somewhat surprising property. Indeed, in our opinion, the most significant finding of this study is the observation that regulatory variation is reduced in both humans and chimpanzees as the cell cultures differentiate from iPSCs to primitive streak. We found a marked overlap between the species in the specific genes that experience this reduction of regulatory variance, indicating high degree of conservation in this process. Before we discuss the potential implications of our observations, we will first discuss a few considerations regarding the iPSC based differentiation models. We argue that the use of iPSC models allows for greater control and transparency of comparative studies in primates, including a better appreciation of caveats that have always affected such studies but were typically cryptic. For more than a decade, comparative genomics in primates has relied on the use of frozen tissues. An implicit assumption underlying the use of these tissues has been that they faithfully reflect interspecies gene regulatory similarities and differences. Yet, we know that gene regulation in these tissues was likely impacted by non-genetic factors, such as the individuals diet, age, and cause of death. In addition, it is nearly impossible to stage frozen tissues with respect to cellular composition. In fact, in practically all studies of

comparative data from frozen tissues, cellular composition was not even measured. Indeed, comparative studies of gene expression in primate frozen tissue samples, including those by our own group, simply assumed that most observed patterns are driven by genetic control. In contrast to frozen tissues, the primate iPSC-based differentiation model allows us to minimize the impact of non-genetic (e.g. environmental) factors. We can also control to a large extent the cellular composition of the samples, and more importantly we can measure and account for differences in cellular composition. Comparative experiments with iPSC-based differentiation are certainly not flawless, but they allow us to more explicitly characterize, and often account for, confounding factors than it was possible with frozen tissue samples. In the case of the current study, we used human and chimpanzee iPSC lines that were generated, and differentiated, using the same protocols. We made considerable efforts to balance the majority of sample processing properties related to our study design with respect to species and time point. For example, the two differentiation batches we used included multiple human and chimpanzee lines (which we also balanced with respect to gender). Admittedly, *in vitro* differentiation protocols are not identical to natural developmental signaling and natural cell-driven developmental processes may be overridden by our administered media conditions. Moreover, although we used an identical differentiation protocol across the entire experiment, we observed a wide distribution of cell purity across samples, with a marked difference between the species in the purity of cultures form days 2 and 3. The fact that we are able to measure and discuss purity and cellular composition as a potential flaw in our study is an advantageous property of the iPSC-based differentiation system. At present, we cannot exclude the possibility that the observed differences in cellular heterogeneity across time points may have driven the observation of reduced regulatory variation exclusively in primitive streak samples. In our opinion, however, this is unlikely, though our arguments are mostly circumstantial and we will not be able to provide a definitive answer without additional single cell experiments (which will be the scope of a future study). Our intuition is

based on the following properties: First, the iPSC cultures are the most homogenous in our experiment and we nevertheless observe a reduction of variation in gene expression levels in day 1 samples. Second, the conclusion of high sharing and similarities between species should be robust (conservative, in fact) with respect to technical differences between species, including in purity. Third, when we account for the purity values measured for the second batch of samples, the relative expression patterns remain similar. That said, to provide definitive answer, we will revisit these questions with an experimental design that involves single cell data from the same comparative differentiation trajectory. Single cell RNA sequencing data will also be able to shed light on our observation that gene regulation in definitive endoderm (day 3), unlike earlier days, does not indicate strong conservation between the species. In definitive endoderm samples, we observed the largest number of DE genes across species, the smallest number of DE genes between two consecutive time points in humans, and the lowest overlap in DE genes between time points. Unlike in the human samples, the chimpanzees had a relatively consistent number of DE genes between any two consecutive time points. Some of these observations might be explained by the cell purity difference between species. It should be noted that the effect sizes for the reduction in variation from day 0 to 1 are small in the chimpanzee samples and that the Pvalue distributions do appear quite different across species. These differences, however, do not invalidate our finding that there is a significant overlap of genes that undergo a reduction of variation in gene expression levels in both species. If the reduction of variation in gene expression levels were spurious within each species, then we would not expect a statistically significant overlap of such genes. The observation of reduced regulatory variation is rather unusual in general, partly due to the unusual design of our study. Indeed, only few comparative studies have been designed to allow one to measure changes in variation over time. One such example, from a completely different context, can be found in a previous study in which monocytes from humans, chimpanzees, and rhesus macaques, which were stimulated with lipopolysaccharide (LPS) to mimic infection

[6]. When comparing gene expression in LPS-stimulated monocytes to that of non-stimulated cells, the authors found a reduction of inter-species variation in gene expression levels in a number of key transcription factors involved in the regulation of *TLR4*-dependent pathways. In our study, we found enrichment for genes in developmental pathways among genes with conserved reduction of expression variation. This observation is consistent with the notion [151] that developmental pathways need to be tightly regulated in general. This notion is also supported by deep conservation and lethality upon disruption seen in many of these genes. Null mutations in mouse orthologs of over 40% of the genes with conserved reduced regulatory variation at primitive streak are associated with embryonic lethality. For example, *Xenopus laevis* embryos with null mutations in the ortholog of human *MIXL1* exhibit abnormalities in primitive streak and node formation. Similarly, homozygous null *MIXL1* mice have abnormal mesoendoderm development and do not survive to birth. *EOMES* homozygous null mice lack trophoectoderm outgrowth and do not properly form the definitive endoderm. This mutation is lethal early in gestation. Overall, regulation of these genes is likely to be finely tuned at early development. Indeed, reduced regulatory variation early in the endoderm differentiation process may be driven by the property of canalization during development. The theory of canalization posits that developmental processes end in a finite number of states despite minor environmental perturbations [127, 151, 150]. Canalization is fundamentally linked to evolutionary states [150], and thus phenotypic robustness; therefore, even when reduced variation in gene expression levels is observed in cell culture, the explanation of canalization is intuitively appealing considering the discrete nature of cell types in an adult animal. Our results suggest that stages subsequent to primitive streak may follow a more relaxed transcriptional regulation with higher influence of individual genotypes.

3.4 Methods

3.4.1 Human and chimpanzee iPSC panels

In this study, we include four chimpanzee iPSC lines (2 males, 2 females) from a previously described panel [42] and six human lines (3 males, 3 females) [30] matched for cell type of origin, reprogramming method, culture conditions and closely matched to passage number (median passage was within 1 passage across species and differentiation batches). We evaluated iPSC lines for pluripotency measures, differentiation potential, lack of integrations and normal karyotypes as described previously [30, 42] (Fig. 3.26-Fig. 3.27). We identified one human individual (H5) that tested positive for episomal vector sequence (Fig. 3.27). This individual was not an obvious outlier in any of our data (Fig. 3.1B- Fig. 3.2), thus we choose to include it in our study. Original chimpanzee fibroblast samples for generation of iPSC lines were obtained from the Yerkes Primate Center under protocol 00612. Human fibroblasts samples for generation of iPSC lines were collected under University of Chicago IRB protocol 110524. Feeder free iPSC 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 [78]. Cell culture was conducted at 37°C, 5% CO_2, and atmospheric O_2.

3.4.2 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 [73] 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. Basal medial for differentiations consisted of 50/50 IMDM/F12

basal media supplemented with 0.5 mg/mL human albumin, 0.7 g/mL Insulin, 15 g/mL holo-Transferrin, 1% v/v chemically defined lipid concentrate, and 450 uM 1-thioglycerol (MTG). For differentiation, basal media was supplemented with the following: day 0 to day 1 (Primitive streak induction) media included 100 ng/mL Activin A, 50 nM PI-103 (PI3K inhibitor), 2 nM CHIR99021 (Wnt agonist), days 1->2 (total of 2 media changes) media included 100 ng/mL Activin A and 250 nM LDN-193189 (BMP inhibitor). Two independent differentiation batches were performed, resulting in replicates for a subset of individuals (3.12). Each chimpanzee was replicated, while only two humans individuals were replicated across the two batches. Replicates were sex-balanced both within and across species. Cell culture was conducted at 37°C, 5% CO_2, and atmospheric O_2.

3.4.3 Purity assessment using flow cytometry

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 timecourse, 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* , which was used at 1/10 of the manufacturer specified concentration (15 ng of each antibody in final volume of 30 μL per staining). We found that the manufacturer recommended dilution produced acceptable results for live cells, however, upon fixing, we observed nonspecific binding by all populations. Thus we determined the optimal antibody titer to maximize the separation between iPSCs (biological negative) and Day 3 definitive endoderm. We found this optimal concentration to be in concordance with that quantity specified by a previous publication using the same antibody clone from a different manufacturer [29]. 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, we used the program FlowJo (http://docs.flowjo.com/d2/credits-2/) to determine scaling. To do so, we used 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 [73]. Day 0 iPSC purity was estimated by dual positive *OCT3/4* and *SOX2* [79] as well as negative staining for *EOMES*. Day 1 primitive streak purity was estimated primarily based on *EOMES* Positive staining [41, 139] but also negative staining for *SOX17*. Day 2 endoderm progenitor purity was quantified by positive staining for *SOX17* expression [149] (*CKIT* could also be used, as its level peaks at day 2) and negative staining for *CXCR4*. Finally, day 3 definitive endoderm purity was estimated by double staining for *CKIT* and *CXCR4* [29]. 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 intensities. 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 is outlined in Fig. 3.12A, with the final purity results for the second batch of differentiation in 3.4. The samples in the first differentiation batch demonstrated hallmarks of improper fixing (highly nonspecific staining of antibodies, most notably for surface markers *CXCR4* and *CKIT*), thus we were unable to determine reliable purity estimates for the first differentiation batch. Purity was also determined using k-means clustering of minimally preprocessed flow cytometry data. After applying the same live, intact, single cell gating scheme used above, we exported compensated fluorescence channel values for processing in R. First, we visually inspected population separation by performing Principal Components Analysis (PCA) on compensated fluorescence channel values. We randomly sampled 1000 cells from each individual at each day, resulting in a total of 4000 randomly sampled cells per individual and 31,000 cells overall (no data was collected for individual H4 at day 3). To assign cells to a specific developmental day, we performed K-means clustering using K = 4 on a correlation matrix representing pairwise correlations between all 31,000 single cells. Four relatively well separated populations were visible after projection onto the first three principal components (Fig. 3.13, top panel) and cluster assignment on the reduced data is shown in Fig. 3.13, middle and bottom panels. We estimated cellular composition at each day by using the PCA to match clusters to a day assignment (Fig. 3.1B).

3.4.4 RNA extraction, library preparation, and sequencing

We collected RNA from iPSCs (day 0) prior to adding day 1 media, and then every 24 hours during the differentiation timecourse for a total of 4 time points representing intermediate cell populations from iPSCs to definitive endoderm (Fig. 3.15B). We extracted the RNA using the ZR-Duet DNA/RNA MiniPrep kit (Zymo) with the addition of an on column DNAse I treatment step prior to RNA elution. We used non-strand specific, polyA capture to generate RNA-seq libraries according to the Illumina TruSeq protocol. To estimate the RNA concentration and quality, we used the Agilent 2100 Bioanalyzer (3.1; Fig. 3.10). We added barcoded adaptors (Illumina TruSeq RNA Sample Preparation Kit v2) and sequenced the 50 base pair single-end RNA-seq libraries on the Illumina HiSeq 4000 at the Functional Genomics Core at University of Chicago on two flowcells (3.2). To minimize the introduction of biases due to batch processing, we chose the RNA extraction batches, library preparation batches, sequencing pools, adaptor names, and flowcells in a manner that maximally partitioned the biological variables of interest (day, species, cell line;3.2 and 3.5). We generated a minimum of 14,424,520 raw reads per sample. We used FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to confirm that the reads were high quality.

3.4.5 Quantifying the number of RNA-seq reads from orthologous genes

We mapped human reads to the hg19 genome and chimpanzee reads to panTro3 using TopHat2 (version 2.0.11) [57], allowing for up to two mismatches in each read. We kept on only reads that mapped uniquely. To prevent biases in expression level estimates due to differences in mRNA transcript size and the relatively poor annotation of the chimpanzee genome, we only kept reads that mapped to a list of orthologous metaexons across 30,030 Ensembl genes available for the hg19 and panTro3 genomes as described previously [13]. Gene expression levels were quantified using the feature counts function in SubRead 1.4.4

[68]. For one sample (C2B at Day 0), the number of raw reads was approximately double the second highest number of raw reads. Therefore, we subsampled the raw reads to approximately the same number of raw reads as the second highest sample. We performed all downstream processing and analysis steps in R (version 3.2.2) unless otherwise stated.

3.4.6 Transformation and normalization of RNA-sequencing reads

After receiving the raw gene counts, we calculated the $\log_2$-transformed counts per million (CPM) for each sample using edgeR [115]. To filter for the lowly expressed genes, we kept only genes with an expression level of $\log_2$ (CPM) > 1.5 in at least 16 samples per species [116]. For the remaining genes, we normalized the original read counts using the weighted trimmed mean of M-values algorithm (TMM) [116] to account for differences in the read counts at the extremes of the distribution and calculated the TMM-normalized $\log_2$-transformed CPM. When we performed principal components analysis (PCA) using the TMM-normalized $\log_2$-transformed log2(CPM) values, we found one outlier (H1B at Day 0, Fig. 3.11A). We removed this sample from the list of original gene counts. We filtered for the lowly expressed genes by retaining genes with an expression level of $\log_2$ (CPM) > 1.5 in at least 15 human samples and at least 16 chimpanzee samples. 10,304 genes remained. We performed TMM-normalization and then performed a cyclic loess normalization with the function normalizeCyclicLoess from the R/Bioconductor package limma [4, 113]. We found that the TMM-normalized $\log_2$ (CPM) values were highly correlated with the TMM- and cyclic loess-normalized $\log_2$ (CPM) values ($r > 0.99$ in the 63 samples; Fig. 3.11B). We used the TMM- and cyclic loess-normalized $\log_2$(CPM) expression values in all downstream analysis unless otherwise stated. We calculated normalized $\log_2$-transformed RPKM values by using the function rpkm with normalized library sizes from the package edgeR [115] (Fig. 3.12C). We measured the "gene lengths" as the sum of the lengths of the orthologous exons and were also used in [42]. This method of calculating RPKM was highly correlated

with a method in which we subtracted $\log_2$ (gene length in kbp) from the TMM- and cyclic loess-normalized $\log_2$ (CPM) values ($r > 0.97$).

3.4.7 Data quality and analysis of technical factors

To assess the data quality, we performed Principal Components Analysis (PCA) on the normalized $\log_2$ (CPM) values from above (Fig. 3.2A). Principal component (PC) 1 was highly associated with day and PC2 was highly associated with species ($r > 0.92$ for each, Fig. 3.2A;3.5). We sought to determine if the studys biological variables of interest were confounded with any of the studys recorded technical aspects (3.1-3.2 and 3.4). First, we calculated which of our 35 recorded technical factors were statistically significant predictors of PCs 1-5 with individual linear models for each technical factor. The 19 statistically significant predictors (FDR cutoff of 10% assessed on the 5x35 matrix) were carried to the second stage. In this stage, we determined which technical factors were associated specifically with either day or species, with individual linear models for each technical factor. We quantified these associations using the *P*values from analysis of variance (ANOVA) for the numerical technical factors and from Chi-squared test (using Monte Carlo simulated *P*values) for the categorical technical factors. Statistical significance was determined by Benjamini-Hochberg adjusted *P*value $< 10\%$ (assessed on the 2x19 matrix). A variable for cell line include a species but not a day component and was tested in this pipeline. They were found to each be confounded with species (B.H. adj. *P*value $< 10^{-4}$) but not day (B.H. adj. *P*value > 0.9), thereby increasing the confidence in our pipeline. We note that when all sequencing pools (mastermixes) were considered together, there was a relationship between adaptor sequence and day (χ^2 test, Benjamini-Hochberg adjusted $P = 0.01$); however, this relationship is substantially weaker when adaptor sequence and day were tested in each of the 4 sequencing pools separately (B.H. adj. $P > 0.9$ in each test). Our most highly dependent variables with day or species were related to properties inherent to the iPSC

model, including harvest density and day (B.H. adj. $P = 0.02$), harvest density and species (B.H. adj. $P = 0.03$), harvest time and day (B.H. adj. $P = 0.01$) (3.2 and 3.5; Fig. 3.14). During this analysis, we observed that the purity estimates were relatively similar across days and between species until the final day (Fig. 3.1B; Fig. 3.12C;3.4). Therefore, it was important to explore how the variance for a given technical factor was partitioned across the biological variables of interest (e.g. across the days, species, and day-by-species interactions). For each recorded technical variable, we created a reduced model and a full model. The reduced model contained only species and day as fixed effects and the technical factor as the response variable. The full model had the same response variable but contained species, day, and a species-by-day interaction as fixed effects. We then compared the two models and reported the significance (3.12). The exact tools used to compare the two models were data-dependent (3.12 and columns 1-2 in 3.12). For numerical data (24 technical factors), we constructed the full and reduced normal general linear models for each technical factor. We compared the models using ANOVA, and extracted the Pvalue directly from ANOVA. For categorical data with 2 levels (3 technical factors), we constructed the two general linear models from the binomial family. We used ANOVA to compare the models and extracted the deviance along with its degrees of freedom. Based on the deviance, we calculated the Chi-Squared statistic and associated Pvalue. 8 technical factors (such as RNA extraction data) contained categorical data with more than two levels. We modeled this data type with multinomial logistic regression with the R/Bioconductor package nnet and used ANOVA to obtain the likelihood ratio statistic and associated Pvalue. We performed this process for each technical factor using data from days 0 and 1 as well as from days 0 to 3 (3.12).

3.4.8 A linear model based framework to perform pairwise differential expression analysis

Differential expression was estimated using a linear model based empirical Bayes method implemented in the R package limma [128, 129]. In order to use a linear modeling approach with RNA-seq read counts, we calculated weights that account for the mean-variance relationship of the count data using the function voom from the limma package [62]. This limma+voom pipeline has previously been shown to perform well with n > 3 biological replicates/condition [111, 133]. For all pairwise differential expression comparisons, the species, day, and a species-by-day interaction were modeled as fixed effects, and individual as a random effect. Individual (cell line) rather than differentiation batch was modeled as a random effect because when using a linear model, individual was most highly correlated with PCs 2 and 3, whereas batch was most highly correlated with PC 10. Since our recorded technical factors were not confounded with our biological variables of interest and did not contribute significantly to the first five principle components of variation (3.5), we did not include any other covariates. We used contrast tests in limma to find genes that were differentially expressed (DE) by species at each day (3.6), DE between days for each species (3.7), and significant day-by-species interactions for days 1-3 (3.9). For each pairwise DE test, we corrected for multiple testing with the Benjamini & Hochberg false discovery rate [10] and genes with an FDR-adjusted *P*values <0.05 were considered DE unless otherwise stated in the text. To find the number of shared DE genes in consecutive time points in each species (Fig. 3.5), we used a two *P*value cutoff system. To be "shared" across species for a given pair of time points (e.g. day 0 to 1), a gene must have an FDR-adjusted *P*value <0.01 in one species and an FDR-adjusted P-value <0.05 in the other species [10]. To estimate the percentage of DE genes in chimpanzees given the observation in humans, we divided the number of genes with an FDR-adjusted *P*value <0.01 in chimpanzees over the number of genes with an FDR-adjusted *P*value <0.05 in humans.

3.4.9 Combining technical replicates

Some analyses did not allow us to model technical replicates explicitly, and treating them as biological replicates would introduce bias in the data. Therefore, we combined technical replicates for the same individual, when available. We calculated the average of the normalized $\log_2$(CPM) values for each cell line at each time point. For day 0, 1 human cell line had a pair of technical replicates that were averaged together. For days 1-3, 2 human cell lines had technical replicates that were averaged. We were able to average technical replicates for each of the four chimpanzee cell lines at each time point. After this process, 6 human data points and 4 chimpanzee data points per day remained, for a total of 40 data points. When we performed principal components analysis (PCA) using these 40 data points, the results were similar to the PCA plot including all the technical replicates (Fig. 3.18A)– PC1 was still correlated with day and PC2 was correlated with species (3.5E). We visually inspected the PCA plot for the distinct clustering of data points with averaged technical replicates and single replicates in the humans, and this potential pattern was not present, increasing our confidence that this process did not introduce bias into the data. We found that the expression values for the 40 samples were robust with respect to the method used to combine the technical replicates. The post-normalization method described above was strongly correlated with a pre-normalization method to combine technical replicates ($r > 0.99$ for the 10,304 genes included in the main analysis; Fig. 3.18B). In our pre-normalization method of combining the technical replicates, we summed the raw counts of technical replicates at each time point (for a total 40 data points) and performed the normalization steps described in the "Transformation and normalization of RNA-sequencing reads" section.

3.4.10 Joint Bayesian analysis with Cormotif

To cluster genes by their temporal gene expression patterns, we used the R/Bioconductor package Cormotif (version 1.22.0), a method that jointly models multiple pairwise differential

expression tests [155]. Unlike other available methods in this class, the Cormotif framework allows for data-set specific differential expression patterns. To identify patterns in expression over time (called "correlation motifs"), expression levels from days 1-3 were compared to those to the previous day for each gene in each species. Since the program does not allow for the explicit modeling of technical replicates (unlike the voom+limma method above), we first ran the program with the expression values averaged across technical replicates. For more information on this process, see the Methods section on "Combining technical replicates". To use Cormotif, we were required to specify the number of correlation motifs to model. We determined a reasonable range by investigating both the Bayesian information criterion (BIC) and Akaike information criterion (AIC). We observed that the BIC and AIC were minimized across many seeds when 7 or 8 correlation motifs were modeled, respectively (3.15; Fig. 3.17A). Thus we further explored models with 7 and 8 correlation motifs. Because Cormotif is not deterministic, we ran Cormotif 100 times and recorded the seed that produced the model with the largest log likelihood (3.15). The best model (the seed with the greatest log likelihood) with 7 correlation motifs is displayed in Fig. 3.17B, and the best model with 8 correlation motifs is featured in Fig. 3.6. We selected the model with 8 correlation motifs to be the primary figure because it had a large log likelihood and all motifs contained more than 100 genes. It should be noted, however, that the two models had very similar correlation motifs. = We were initially conservative when assigning a gene to a specific correlation motif. Following the advice of the Cormotif authors [155], a gene must have a posterior likelihood estimate of 0.5 to be called DE between time points and < 0.5 to be considered not DE (3.10A). We also used this assignment criteria when using Cormotif to compare expression levels using different combination methods (Fig. 3.17C) and to compare all time points to day 0 (Fig. 3.17D). For a trajectory to be defined as DE, the trajectory in humans and chimpanzees needed similar posterior probabilities of differential expression ($\leq$ 0.20) at each comparison along the trajectory. Using topGO [2], we tested for enrichment

of Gene Ontology (GO) biological processes enrichment analysis on various combinations of correlation motifs (3.10B-D). To test for significance, we used the same parameters as [16]. This included the use of Fishers Exact Test, with topGOs weight01 algorithm to account for the correlation among GO categories in its graph structure. Categories with Pvalue < 0.01 were considered significant. To determine the categories enriched in Motif 4+7 group only, we ensured that these categories were not significant in any other group ($P< 0.01$; 3.10B-D). To test this in a more rigorous way, we then compared the enriched categories for Motif 4+7 group given the genes in the two other groups (3.10E) using the compareCluster function in the R package clusterProfiler. We used a q value cutoff of 0.05 [10] and set the size of genes annotated by Ontology term for testing parameter between 3 and 3000.

3.4.11 Global analysis of variation in gene expression levels

We calculated the variance in gene expression level for each gene in each species. Since the largest theoretical range of a variance is from 0 to infinity, we performed a $\log_2$ transformation to each variance value (Fig. 3.7). For the analysis with LCLs and the 4 tissues, we used normalized gene expression data from the GTEx Portal (release V6p; gene expression data is in RPKM) from lymphoblastoid cell lines and 4 additional tissues [1]. Three of the four tissues are derived from the endoderm germ layerliver, lung, and pancreas. The heart tissue is mesoderm-derived, for comparison. We identified all of the genes that were expressed and passed GTExs filtering criteria (n = 17,542 genes) as well as the individuals that had contributed samples to all 5 tissues (n = 6). We then calculated and plotted the $\log_2$ variance of the normalized gene expression using the same pipeline as before (Fig. 3.19A). For samples with associated purity values (n = 30), we calculated the effect of purity gene expression levels. We regressed out effect of purity on gene expression levels on a gene-by-gene basis using the linear model function in R in each species independently. We then calculated the $\log_2$ variance of these residuals for each gene. We shifted the $\log_2$ variances for these residuals

so that the median at day 0 for each species = 1. For samples without purity values (n = 32), we calculated the $\log_2$ variance of the gene expression levels for each gene, and scaled the values so that the median at day 0 for each species = 1. For each gene at each day for each species, we averaged the scaled $\log_2$ variances of the residuals and the $\log_2$ variances of the gene expression levels (Fig. 3.19B). We then performed a one-sided t-test between the distribution of $\log_2$ (variances in gene expression) from day 0 and day 1 in each species, with the alternative hypothesis that the variation was greater in day 0 than day 1. We compared the effect sizes of interspecies DE genes with a one-sided MannWhitney U test on magnitudes of effect sizes (Fig. 3.19C). We tested the null that there was no change in $\log_2$ fold change in gene expression across the species from day 0 to day 1, with the alternative hypothesis that the average magnitude of effect of DE genes (FDR = 5%) was greater in day 0 than day 1.

3.4.12 Gene-by-gene analysis of variation in gene expression levels and calculating the proportion of true positives

To determine if there was an enrichment of genes undergoing changes in variation one species, we used an F test to compare two variances in R (var.test command) for each gene using the averaged $\log_2$(CPM) expression values of technical replicates. In these tests, the null hypothesis was no change of variance in the gene expression levels between days and the alternative hypothesis was a reduction in variation of gene expression levels between two time points (a one-sided test). We calculated the Pvalues for the F statistics from each test and plotted the densities using ggplot2 [157]. If a Pvalue distribution appeared to be even slightly skewed towards small Pvalues, we used the R package qvalue to determine $\hat{\pi_0}$,the true proportion of null statistics from a given Pvalue distribution [135]. Its complement, $\hat{\pi_1}$, is considered the proportion of significant tests from a Pvalue distribution. We used this process to analyze the reduction in variation in each species from days 0 to 1, 1 to 2,

2 to 3, and 0 to 2 (Fig. 3.8A-B). Afterwards, we used the same procedure (F tests) to test the alternative hypothesis that the variation of gene expression increased between two time points. We determined $\hat{\pi}_0$, and $\hat{\pi}_1$,in the same manner as above to analyze the increase in variation in each species from days 0 to 1, 1 to 2, 2 to 3, and 0 to 2 (Fig. 3.8A-B).

3.4.13 Estimating the proportion of genes that undergo a change in variation in both species

We then estimated the true proportion of significant genes shared across species for a given set of time points. Rather than take the intersection of the significant genes (for which we would be underpowered), we adopted a method from Storey and Tibshirani 2003 (Storeys $\hat{\pi}_0$,) [135]. This method was recently implemented by Banovich 2016 to determine the sharing of quantitative trait loci (QTLs) from different cell types [5, 135]. Using the Pvalue distributions generated in the previous section, we subset the genes in species 2 conditioned on its F statistic significance in species 1 (unadjusted $P\text{value} < 0.05$). To test for an enrichment of small Pvalues, we used the Pvalues from species 2 to determine $\hat{\pi}_0$, using the same process as the previous section (Fig. 3.9AC-F, Fig. 3.22). We then repeated this process for other Pvalue cutoffs, including 0.01 and 0.10 (Fig. 3.24 and Fig. 3.25). To determine robustness with respect to the number of genes considered significant in species 1, we calculated $\hat{\pi}_0$, for species 2 conditioned on 100 genes with the lowest Pvalues in species 1. This process was repeated for the top 101 to all 10,304 genes (Fig. 3.23).

3.4.14 Estimating the null hypothesis for the proportion of genes that undergo a change in variation in both species

We then estimated the true proportion of significant genes shared across species for a given set of time points. Rather than take the intersection of the significant genes (for which we

would be underpowered), we adopted a method from Storey and Tibshirani 2003 (Storeys $\hat{\pi_0}$,) [135]. This method was recently implemented by Banovich 2016 to determine the sharing of quantitative trait loci (QTLs) from different cell types [5, 135]. Using the Pvalue distributions generated in the previous section, we subset the genes in species 2 conditioned on its F statistic significance in species 1 (unadjusted Pvalue < 0.05). To test for an enrichment of small Pvalues, we used the Pvalues from species 2 to determine $\hat{\pi_0}$, using the same process as the previous section (Fig. 3.9AC-F, Fig. 3.22). We then repeated this process for other Pvalue cutoffs, including 0.01 and 0.10 (Fig. 3.24 and Fig. 3.25). To determine robustness with respect to the number of genes considered significant in species 1, we calculated $\hat{\pi_0}$, for species 2 conditioned on 100 genes with the lowest Pvalues in species 1. This process was repeated for the top 101 to all 10,304 genes (Fig. 3.23).

Estimating the null hypothesis for the proportion of genes that undergo a change in variation in both species To determine the null hypothesis for the $\hat{\pi_1}$, based on conditioning, we performed permutation tests. First, we combined the unadjusted Pvalues from the F test for a reduction in variation from days 0 to 1 in chimpanzees (species 1) and humans (species 2). We used the randomizeMatrix function in the R package picante [56] to permute the Pvalues of species 1 and then merged this Pvalue distribution with the Pvalue distribution from species 2. We then determined$\hat{\pi_0}$ in species 1 conditioned on its Pvalue significance in species 2 (unadjusted Pvalue < 0.05). We repeated this process a total of 100,000 times and found the complement of the 100,000 values (3.14). We defined the permuted null hypothesis as the mean $\hat{\pi_1}$, value. We then repeated this process, with humans as species 1 and chimpanzees as species 2. For the enrichment analysis, we classified the genes with a reduced variation of gene expression in both species using a 2 Pvalue cutoff in the results of the F tests ($P< 0.05$ in one species and $P< 0.10$ in the other species). We chose this method because names of the overlapping genes can not be determined by Storeys approach. As described in the Methods section, "Joint Bayesian Analysis with Cormotif", we performed the

enrichment analysis using topGO with a significance of $P< 0.01$ [2]. To determine the genes associated with embryonic lethality response to perturbation, we entered our list of genes into the Mouse Genome Database at the Mouse Genome Informatics website provided by the Jackson Laboratory (http://www.informatics.jax.org/batch) and selected the "Mammalian Phenotype" option. With this output (3.13D-E), we eliminated any feature type that was not a protein coding gene and those genes without an associated phenotype. For each gene list, we calculated how many genes contained at least one of the following terms: "embryonic lethality", "prenatal lethality" or "lethality throughout fetal growth and development".

3.5 Supplementary Information

3.5.1 Supplementary Figures

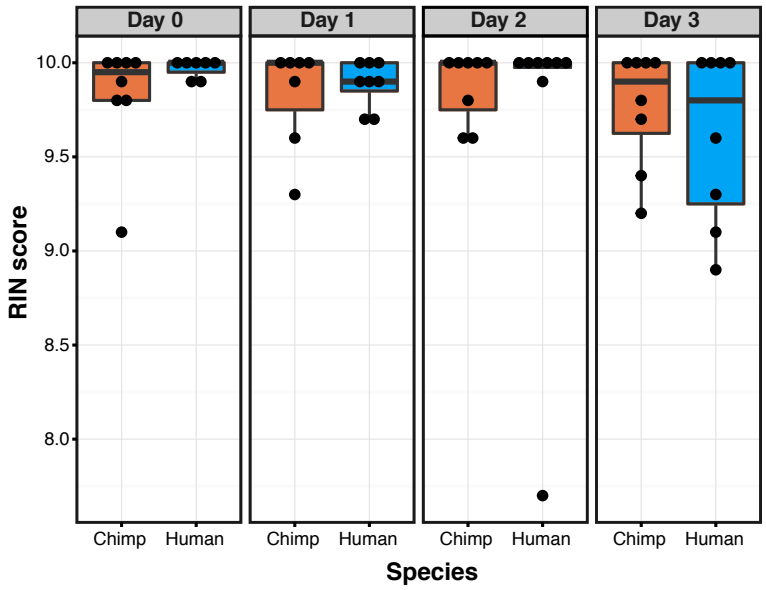


Figure 3.10: **RNA Integrity Number (RIN) scores across biological variables of interest.** RNA quality was not confounded with day, species, or batch (Benjamini-Hochberg adjusted Pvalue > 0.10 for individual linear models). Each point represents one sample.

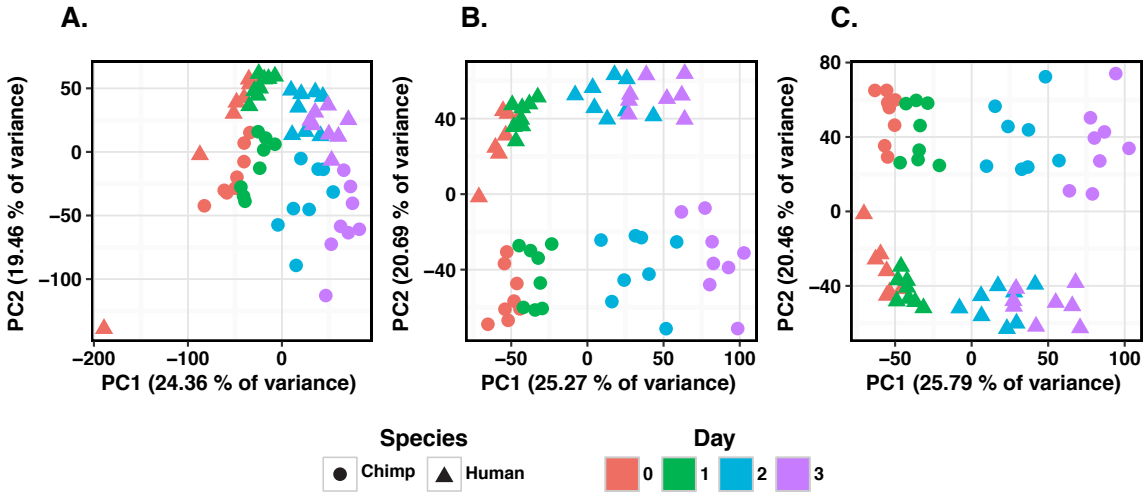
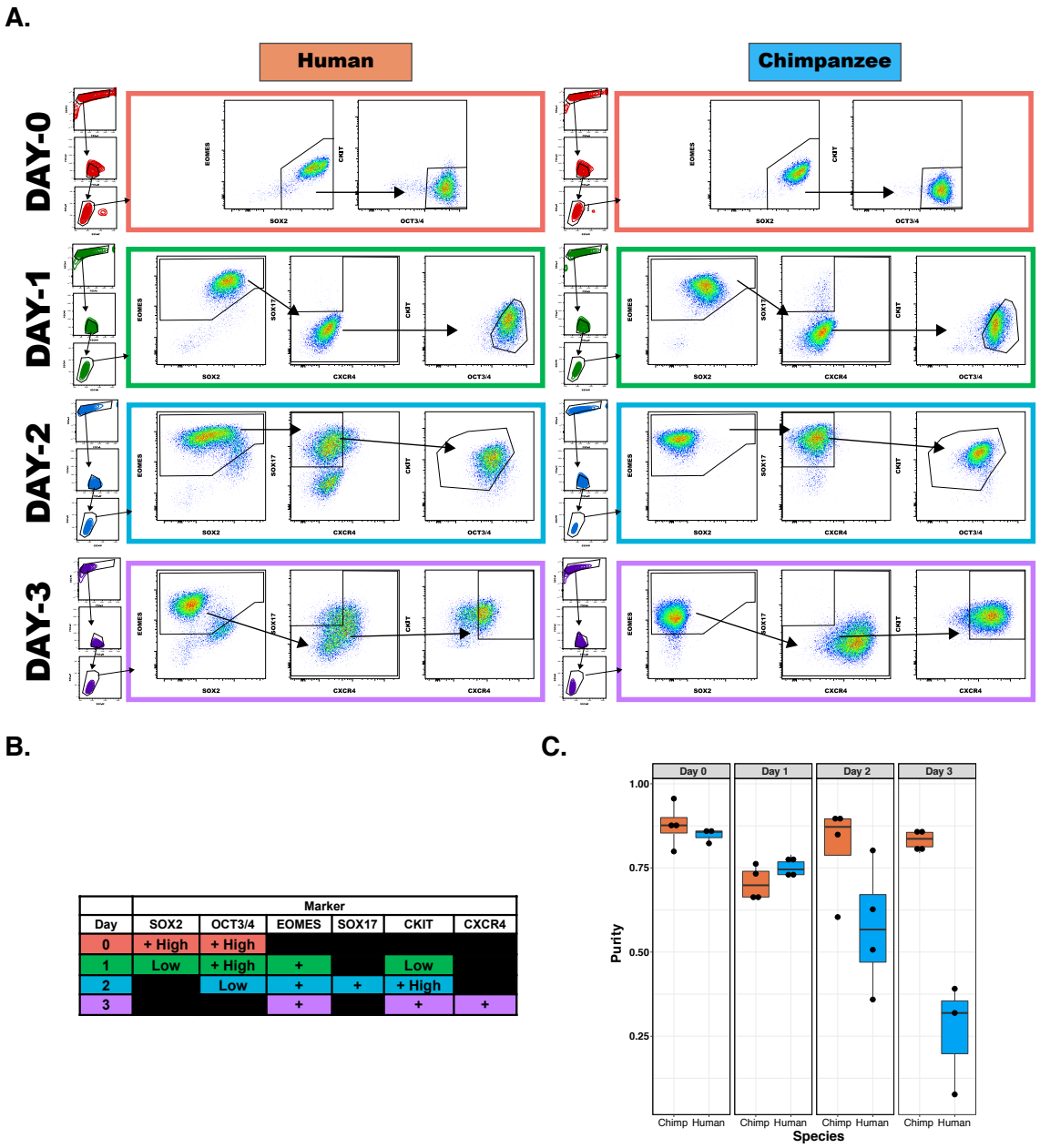


Figure 3.11: **: Principal components analysis (PCA) of normalized data.** A. Prior to the removal of one outlier sample (H1B at day 0). PC1 separates the data from H1B at day 0 from all other samples. B. PCA of the TMM-normalized $\log_2$-transformed CPM. PC1 separates the samples by time point (day). PC2 separates the human and chimpanzee samples. C. PCA of the library-size normalized $\log_2$ reads per kilobase of transcription per million mapped reads (RPKM). PC1 is strongly correlated with cell state (day). PC2 is strongly associated with species.



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

Figure 3.12: **Purity gating information.** A. Gating example for human (H5B) and chimpanzee (C4B) timecourses. Dead cells, debris and doublets were removed prior to analysis of fluorescent markers using SSC and FSC gates as shown to the left of each panel. B. Summary of purity gating scheme. Minimal florescence is expected for the markers in black. C. Box plot of purity estimates by day and species.

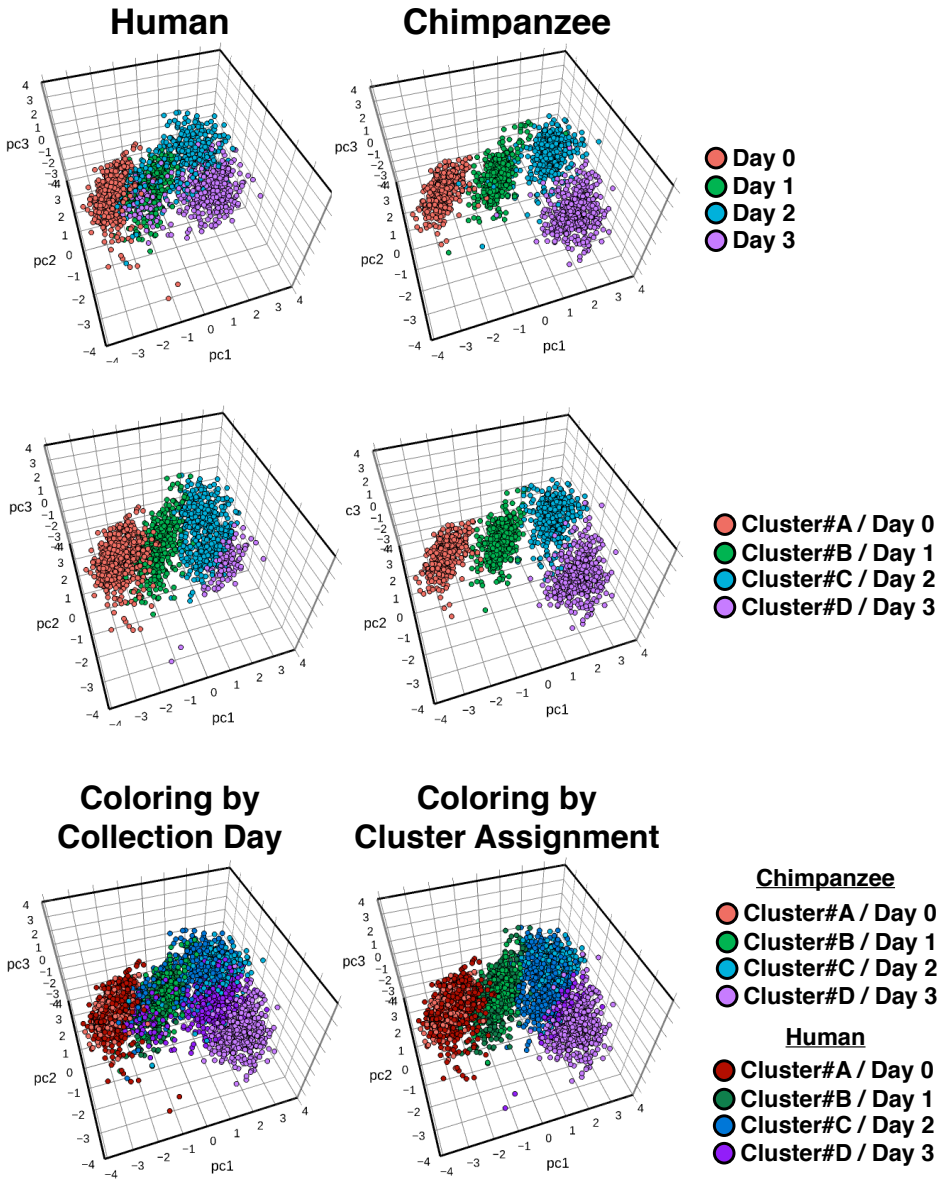


Figure 3.13: **Clustering of fluorescence values.** Compensated fluorescence values projected onto the axes of the first three principal components. For plotting, we randomly sampled 400 cells from each individual (total of 1600 cells per species) from each day separately in humans (left panel) and chimpanzee (right panel). Thus each plot is a composite of all days and all individuals, for a given species. The same set of 1600 cells is used throughout the three panels and only the coloring indicating harvest day or cluster assignment is changed.

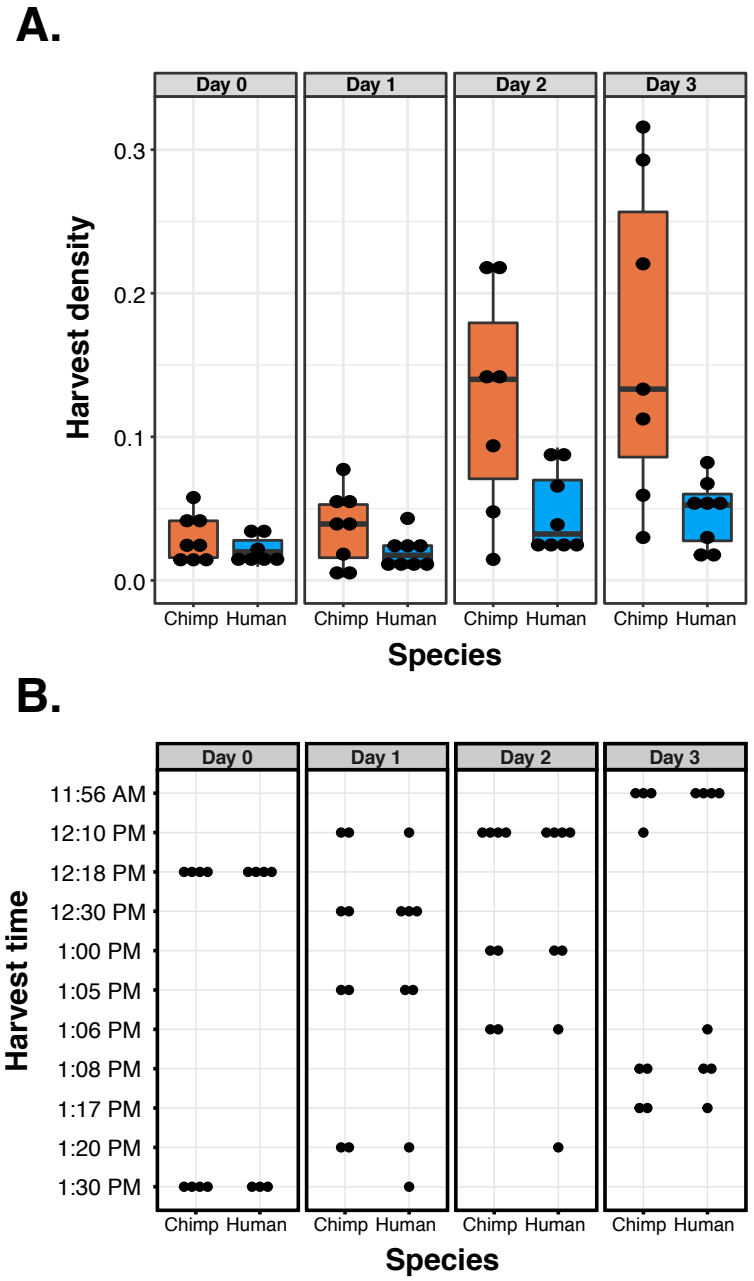


Figure 3.14: **Distribution of potential confounder variables by day and species.** A. Harvest densities for each sample by day and species. B. Harvest times for each sample by day and species. For more information about harvest densities or times.

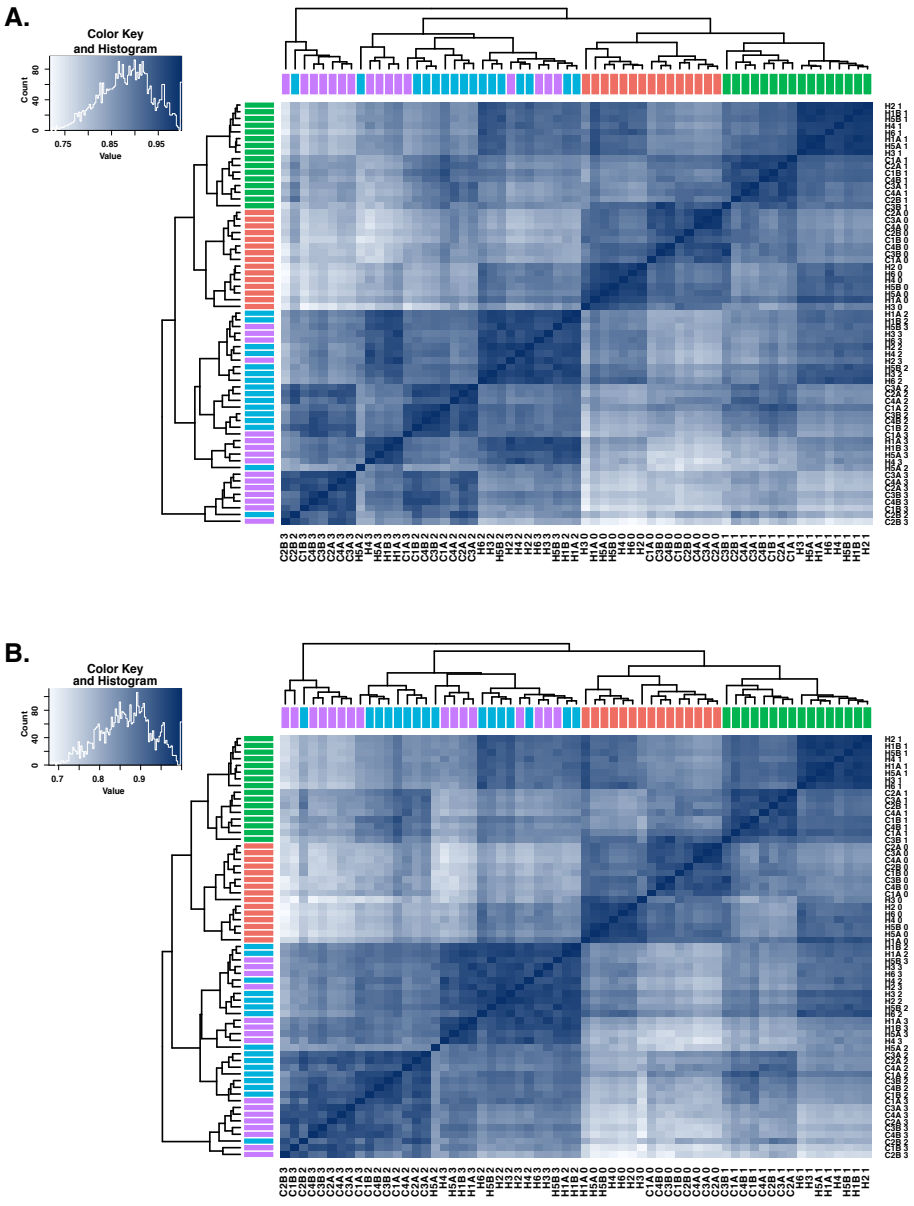


Figure 3.15: **Correlation matrices of normalized $\log_2$(CPM) gene expression values from 10,304 genes.** A. Heat map of the Spearmans correlation matrix of normalized $\log_2$(CPM) gene expression values from 10,304 genes. Each square represents the Spearmans correlation of the normalized expression values between two samples. Samples from the same day tend to be more highly correlated (dark blue) than samples from different differentiation days (light blue). B. Heat map of the Pearsons correlation matrix of normalized gene expression values from 10,304 genes. The results are similar to A.

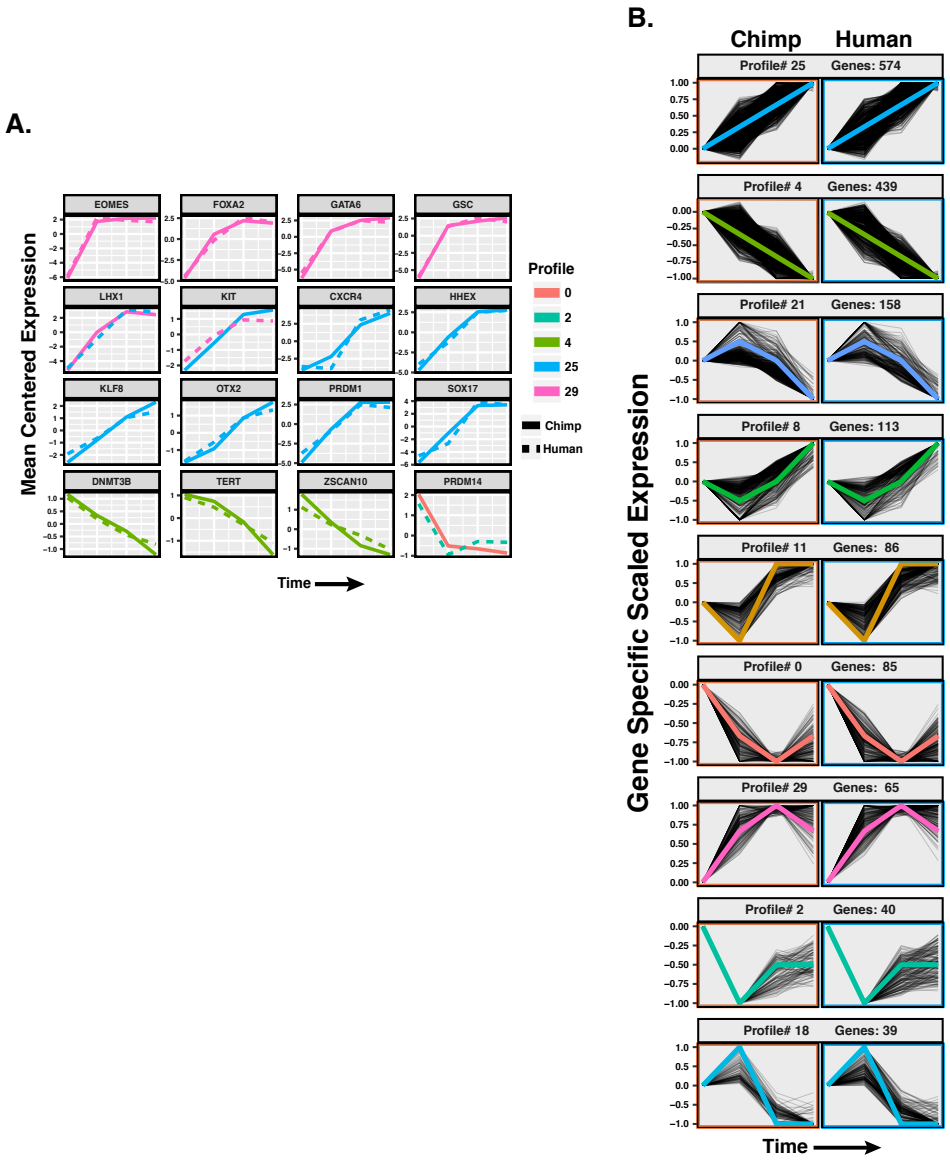


Figure 3.16: **Classifying genes into temporal profiles with Short time course expression miner (STEM) based on TMM-normalized $\log_2$(CPM) expression values.** A. Temporal profiles for 16 known critical regulators or markers of the differentiation process. Chimpanzee time course is shown using a solid line while the human time course is shown using a dashed line. Expression is averaged and scaled by mean centering in order to demonstrate magnitude differences at higher resolution. B. Temporal profiles determined significant for either chimpanzee or human using STEM. Only genes that were assigned to the same cluster in both species were plotted (grey lines). The colored lines indicate the canonical model profile as defined by STEM. Colors indicate cluster assignment and are the same as used in A. Expression across the time course is averaged and scaled (constraining the data to fall between 1 and -1 by dividing by the absolute value of the most extreme value observed throughout the time course) for each gene.

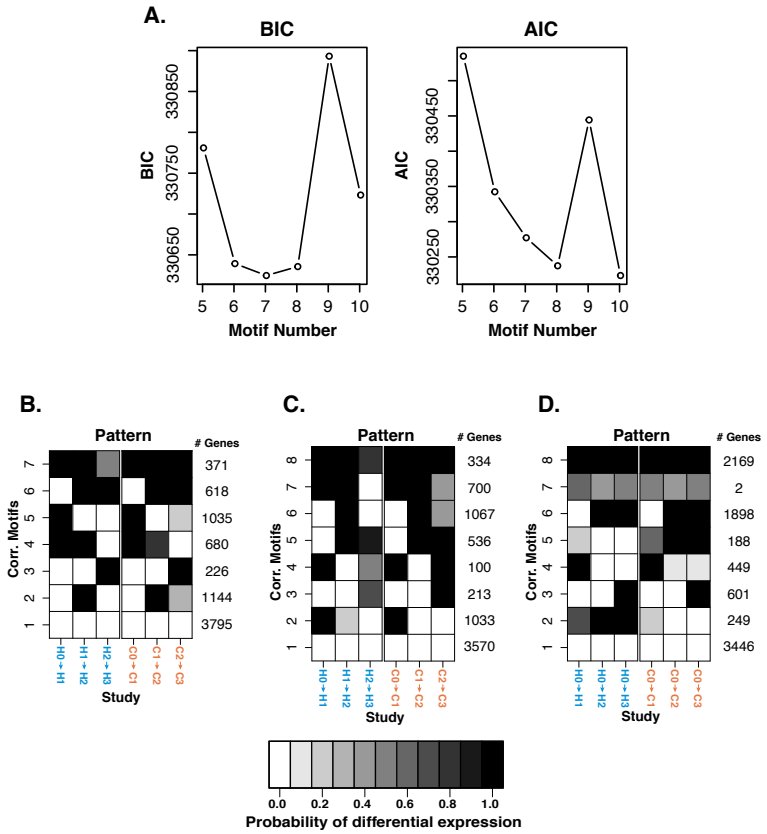


Figure 3.17: **Assessing the robustness of Cormotif results.** The shading of each box represents the posterior probability that a gene is DE between two time points in a given species. Each row ("correlation motif") represents the most prevalent expression patterns along the trajectory. The Bayesian information criterion (BIC) and Akaike information criterion (AIC) are goodness-of-fit measurements for the number of motifs in a model (for a given seed). A. The BIC and AIC for models from seed 12345 allowing for 5 to 10 correlation motifs. These plots highlight that the BIC and/or AIC was often minimized in models with 7 and 8 correlation motifs. B. The predominant expression patterns from a model with 7 correlation motifs (seed 4, see Methods for information about seed selection). The degree of conservation across species calculated using this model is slightly higher than the estimate from Figure 5 (80%, see Supplementary Information). C. The 8 correlation motifs and number of genes assigned to the correlation motifs when using a pre-normalization method to combine technical replicates (seed 66). The predominant expression patterns and the number of genes assigned to each expression patterns are similar to Figure 5. D. The 8 correlation motifs and number of genes assigned to the correlation motifs when the data points are compared to day 0 rather than the previous day (seed 66). Here, we also observed similar expression trajectories across species (particularly motifs 1, 3, 4, 6, 7, 8).

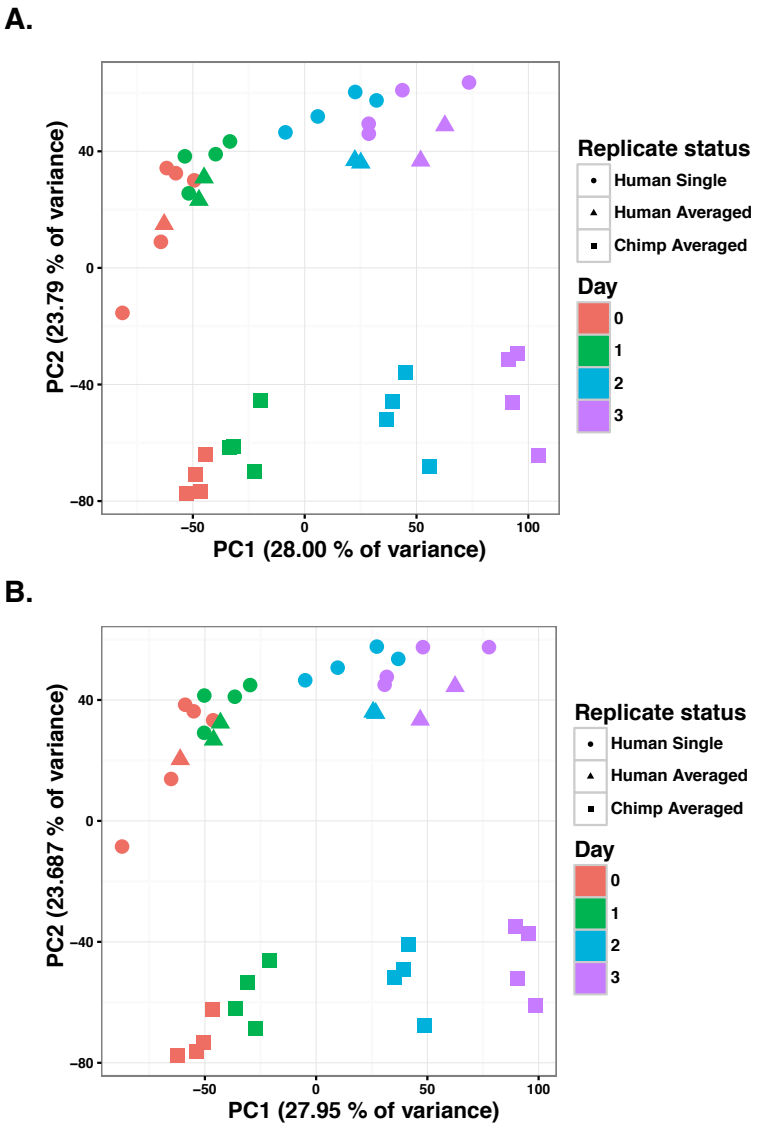


Figure 3.18: **Principal component analysis (PCA) of normalized, combined data (n = 40)** A. PCA of the normalized data after averaging the $\log_2$(CPM) gene expression values across technical replicates (n = 40 data points). PC1 is highly correlated with time point (day) and PC2 is highly correlated with species. B. PCA of the normalized data after combining the gene counts of technical replicates prior to normalization (n = 40 data points). Note that this plot produces similar results to A.

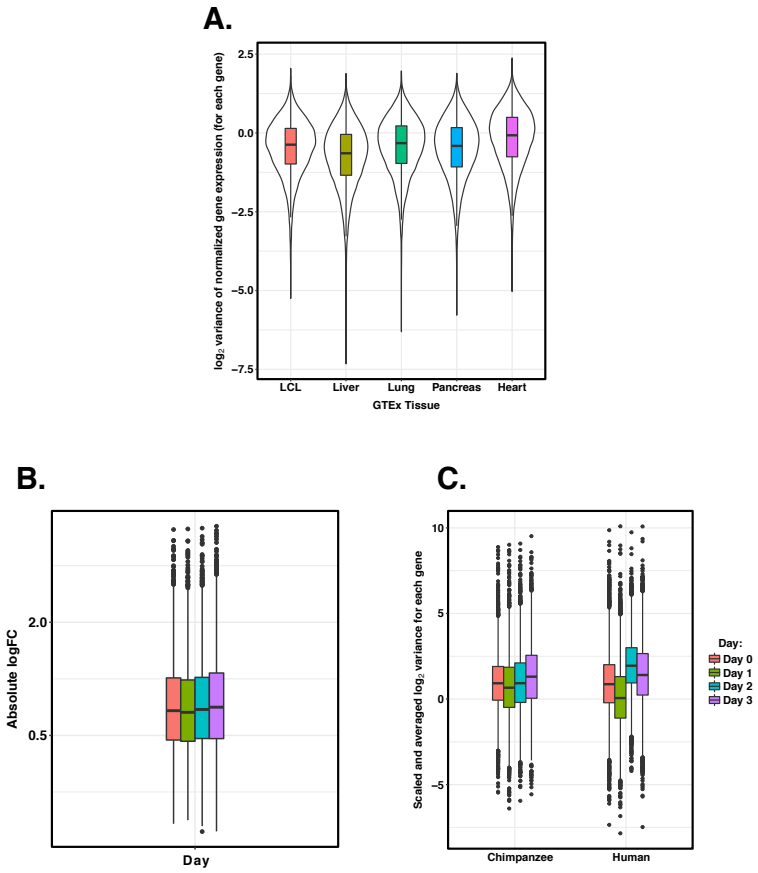


Figure 3.19: **Assessing the variance of gene expression levels.** A. Boxplot of the $\log_2$ variances of normalized RPKM gene expression levels from lymphoblastoid cell lines (LCLs) and 4 tissues collected by the GTEx Consortium. The distribution of variance of gene expression levels in LCLs are similar to those of the 4 GTEx tissues. B. Box plot of the scaled and averaged $\log_2$ variances of gene expression levels for each gene. For samples with associated purity values (n = 30), we regressed out the effect of purity and then calculated the $\log_2$ variance of the residuals for each gene. We shifted the $\log_2$ variances of these residuals (n = 30) to match the scales of the $\log_2$ variances of gene expression levels in samples without purity values (n = 32) independently and then averaged the values for each gene. The $\log_2$ variance values are lowest at the primitive streak (day 1). C. Distributions of effect sizes of differentially expressed (DE) genes between species across days. Box plot of the absolute value of the log fold change (Absolute logFC) in gene expression between species, at each day. The effect sizes are lowest at the primitive streak (day 1), though the overall effect size is quite modest.

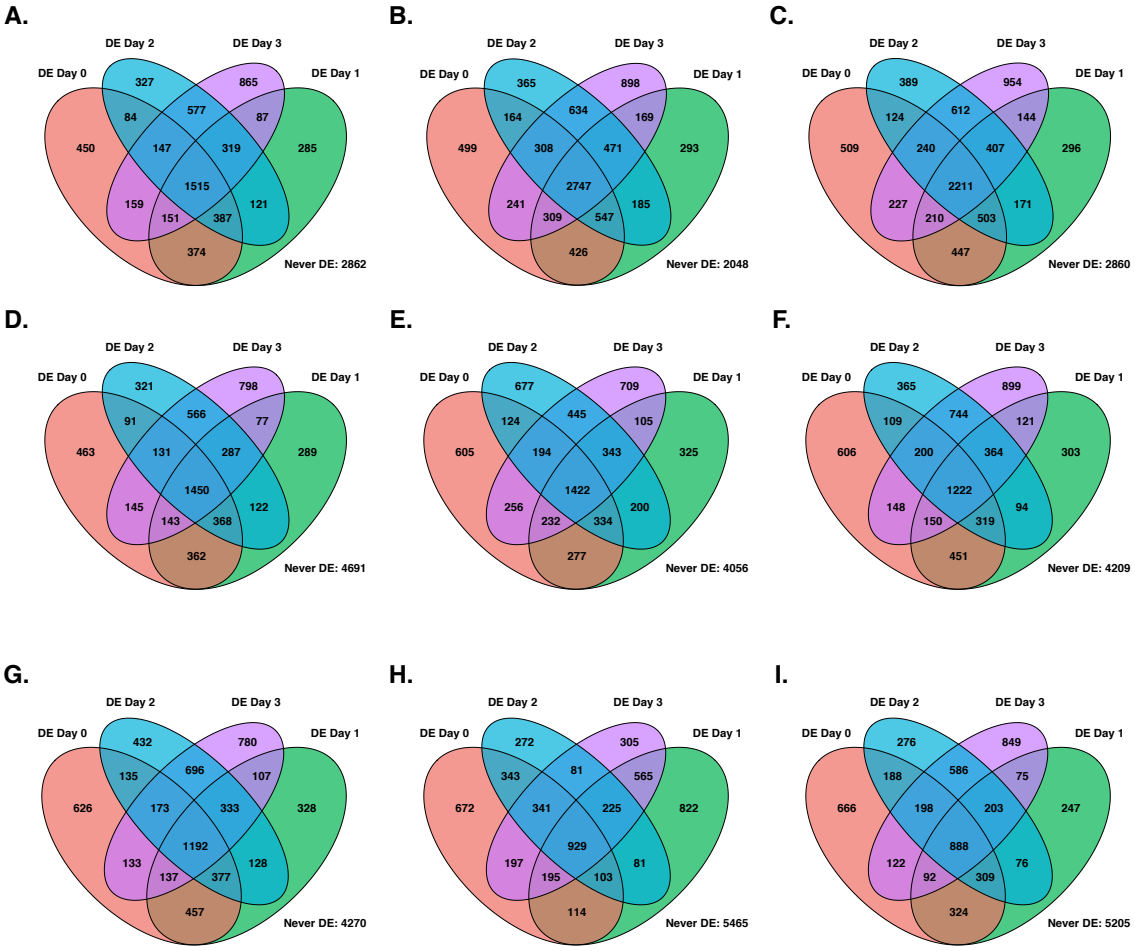


Figure 3.20: **The number of interspecies DE genes using a pairwise, linear model-based approach (limma) is lowest in the primitive streak state.** A. Venn diagram of all DE genes across species at 1% FDR. B. Venn diagram of all DE genes across species at 10% FDR. C. Venn diagram of all DE genes across species at 5% FDR using normalized RPKM values. D. Venn diagram of all DE genes across species at 5% FDR with the samples from differentiation batch 1 (n = 31). E. Venn diagram of all DE genes across species at 5% FDR with the samples from differentiation batch 2 (n = 32). F. Venn diagram of all DE genes across species at 5% FDR with only the samples for which we determined purity (n = 30). G. Venn diagram of all DE genes across species at 5% FDR with 24 samples. Each comparison group contained the samples with the 3 highest purity estimates pairs (e.g. 3 human samples at day 0, 3 chimpanzee samples at day 0, etc.). H. Venn diagram of all DE genes across species at 5% FDR with 24 samples. These samples had the 3 lowest purity estimates for each of the 8 day-species pairs. I. Venn diagram of all DE genes across species with a global correction at 5% FDR. This global correction approach combines the *P*values from all contrasts into one vector (our study contained 13 contrasts) and then applies a *P*value correction on this single vector.

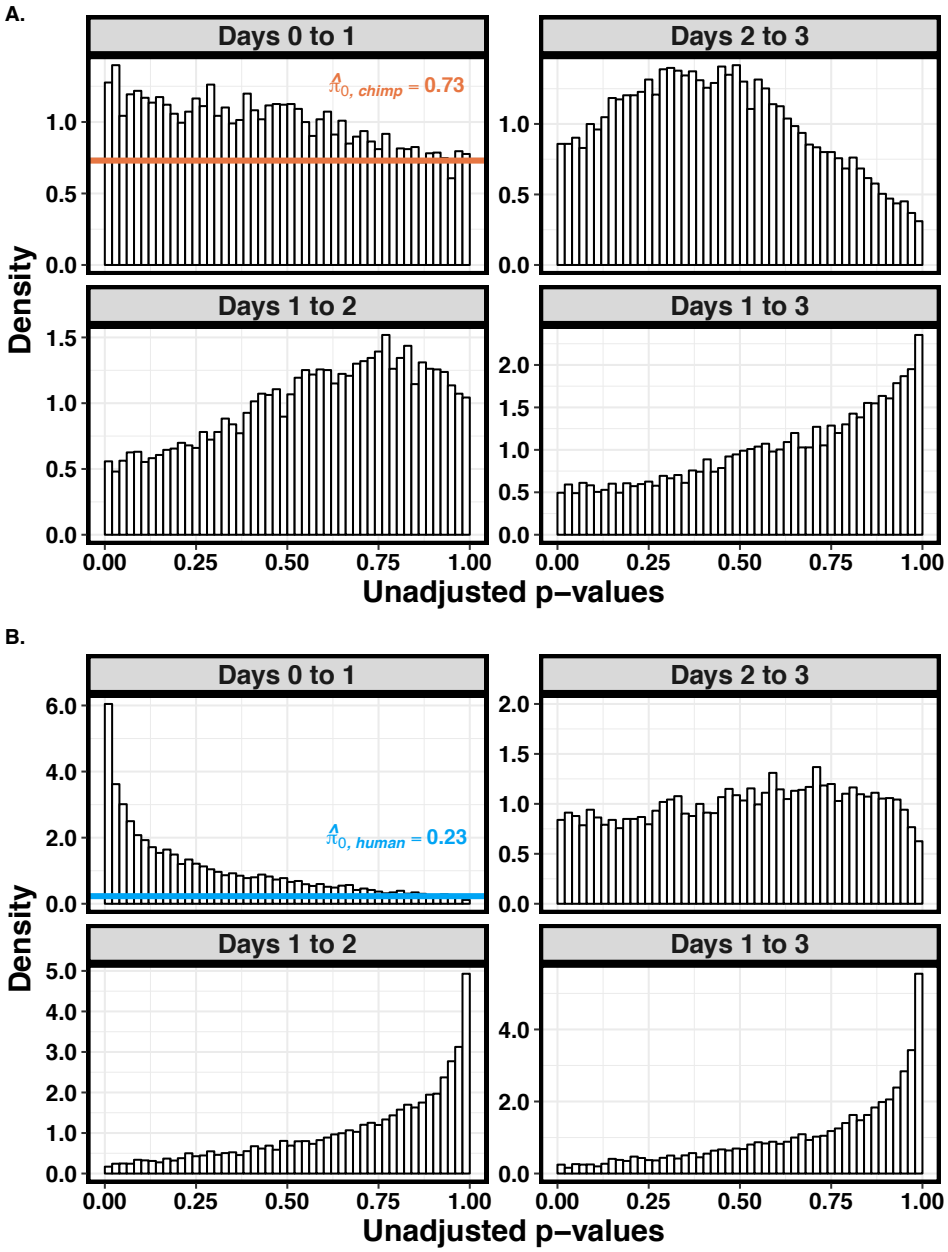


Figure 3.21: **Reduced variation in gene expression at the primitive streak is maintained when using a bootstrap method to calculate $\hat{\pi_0}$** Results from F tests against the null hypothesis that there was no reduction in variation in gene expression levels (day 1 versus 0, 2 versus 1, 3 versus 2, and 3 versus 1) in the chimpanzee (A) and human (B) samples. In both plots, the largest proportion of genes with reduced variation occurred at the primitive streak state.

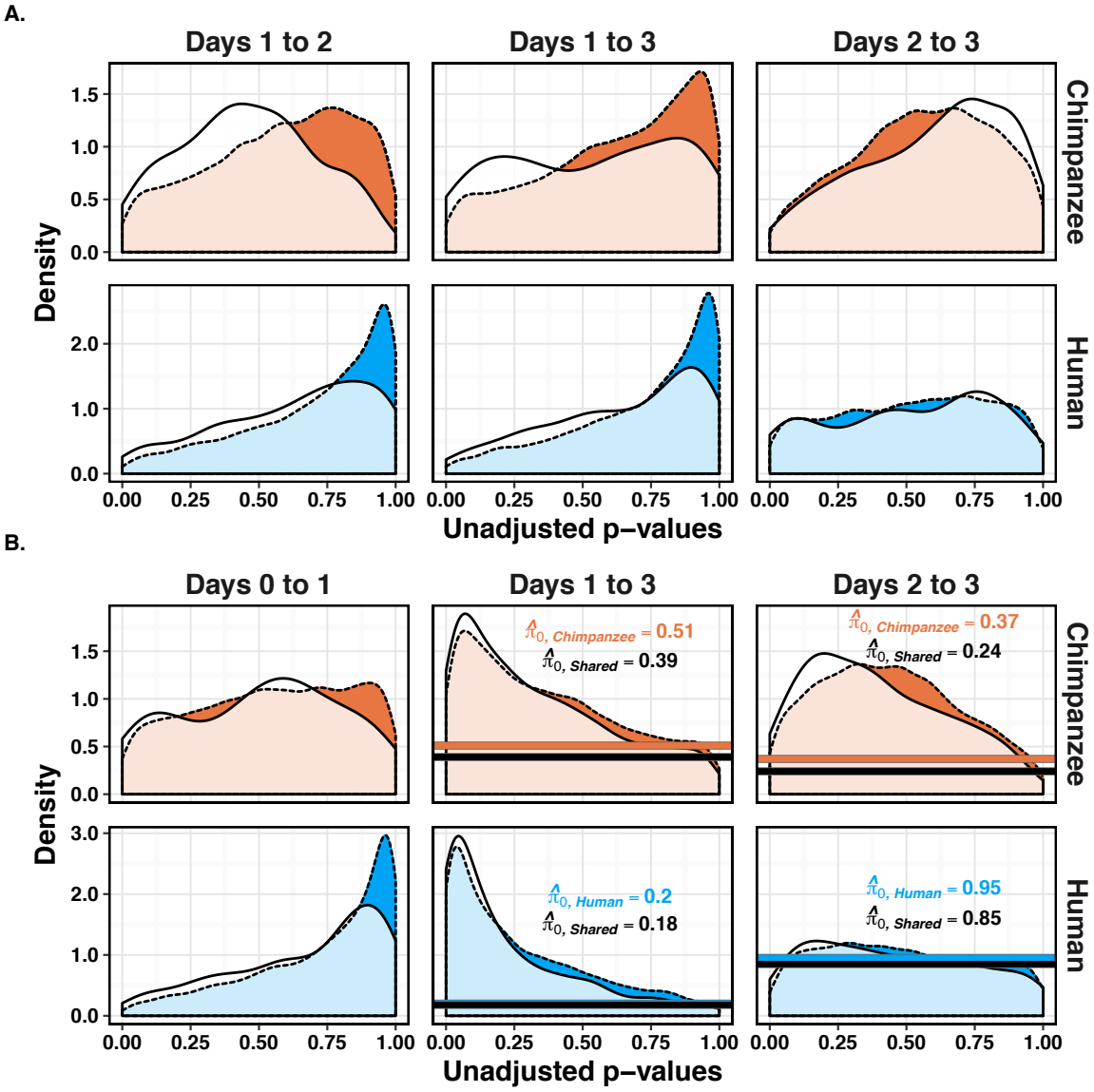


Figure 3.22: **The reduction of variation in gene expression is localized to the primitive streak.** A. *P*value distributions from F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course. Each plot shows the *P*value distribution from F tests for all genes from a given species (orange for chimpanzee and blue for human) and from only genes in that species for which reduced variation was detected in the other species ($P< 0.05$, white). B. Same as (A) except the results from F tests against the null hypothesis that there was no increase in gene expression levels (increased in day 1 compared to 0, 3 compared to 1, 3 compared to 2).

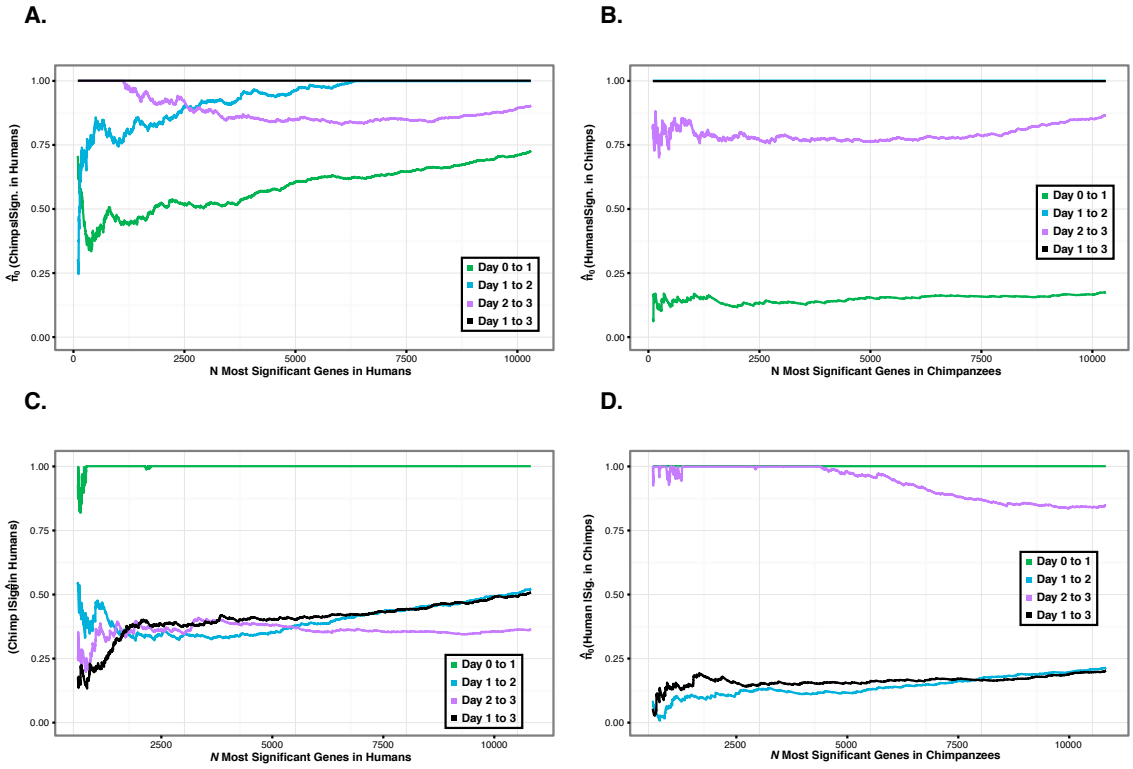


Figure 3.23: **The patterns of change of variation in gene expression are robust with respect to a cutoff based on the number of genes.** *P*value distributions of F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course in human (A) and chimpanzee (B) samples. In each test, Only the N genes whose variation was classified as reduced between states in other species were included. $\hat{\pi}_0$. is the estimated proportion of null tests in each distribution. We then plotted the *P*value distributions of F tests against the null hypothesis that there was no increase in variation in gene expression levels as the samples progress along the time course in human (C) and chimpanzee (D) samples. Again, only the N genes whose variation was classified as reduced between states in other species were included.

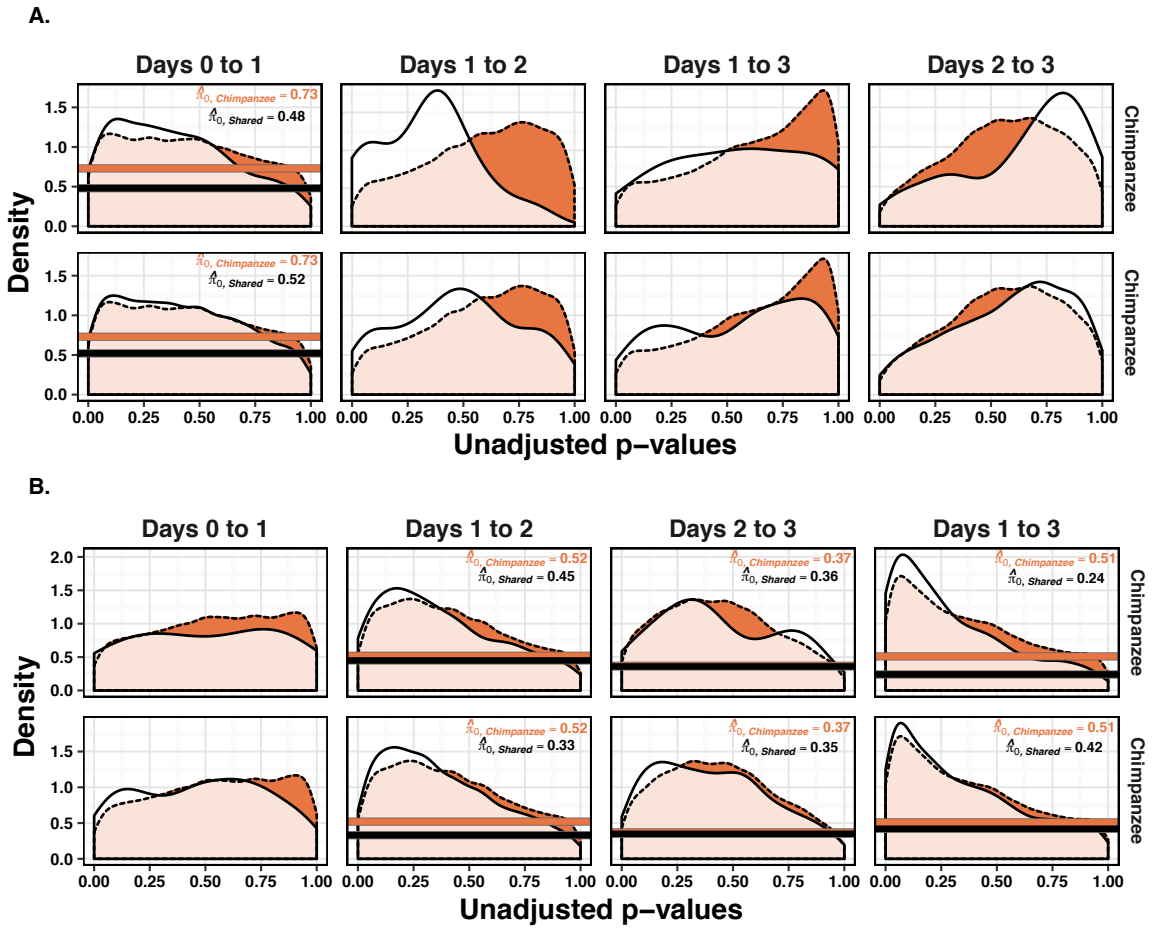


Figure 3.24: **The patterns of change of variation in gene expression are robust with respect to *P*value cutoff in the chimpanzee samples.** $\hat{\pi}_0$ is the estimated proportion of null tests in each distribution. A. *P*value distributions from F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course for all genes from the chimpanzee samples (orange) and the genes in chimpanzees that were significant in the human samples (white, $P< 0.01$ on top, $P< 0.1$ on bottom row). $\hat{\pi}_0$. in the Days 1 to 2 test could not be computed in the top row. B. *P*value distributions of F tests against the null hypothesis that there was no increase in variation in gene expression levels as the samples progress along the time course for all genes from the chimpanzee samples (orange) and the genes in chimpanzees that were significant in the human samples (white, $P< 0.01$ on top, $P< 0.1$ on bottom).

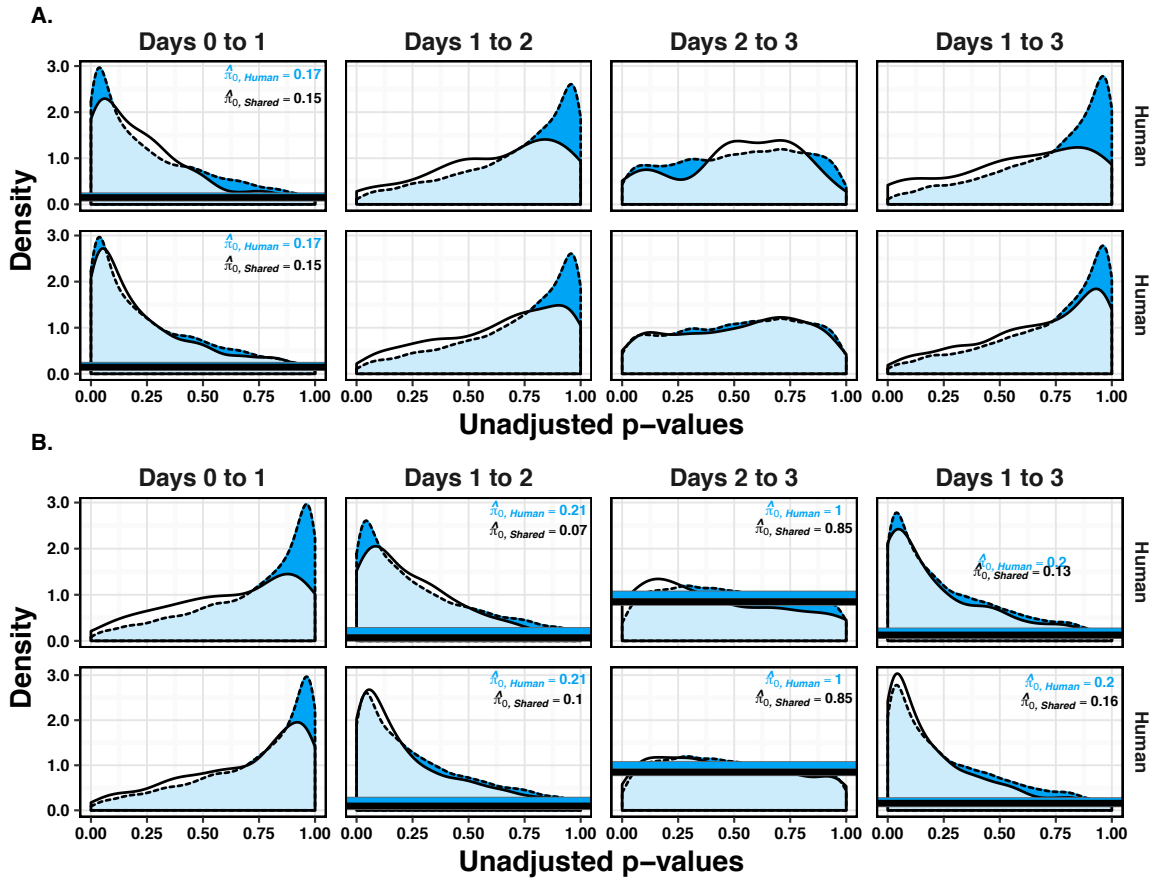


Figure 3.25: **The patterns of change of variation in gene expression are robust with respect to *P*value cutoff in the human samples.** $\hat{\pi_0}$ is the estimated proportion of null tests in each distribution. A. We plotted the Pvalue distributions of F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course for all genes from the human samples (blue) and the genes in humans that were significant in the chimpanzee samples (white, $P< 0.01$ on top, $P< 0.1$ on bottom row). B. Next, we plotted the Pvalue distributions of F tests against the null hypothesis that there was no increase in variation in gene expression levels as the samples progress along the time course for all genes from the human samples (blue) and the genes in humans that were significant in the chimpanzee samples (white, $P< 0.01$ on top, $P< 0.1$ on bottom).

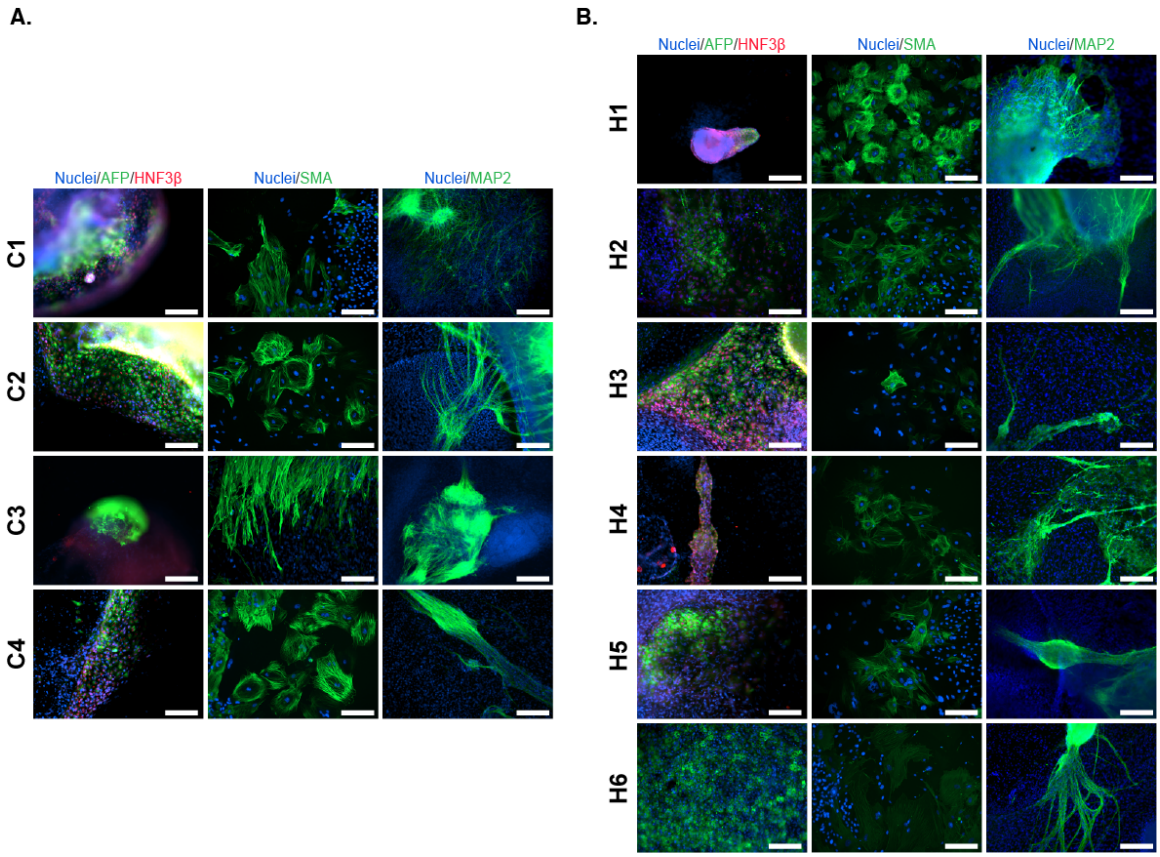


Figure 3.26: **Assay of pluripotency for iPSC lines used in this study.** A. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for chimpanzee iPSC lines. B. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for human iPSC lines, antibodies identifying celltypes derived from the three germ layers as indicated.Scale bar: 200 μM.

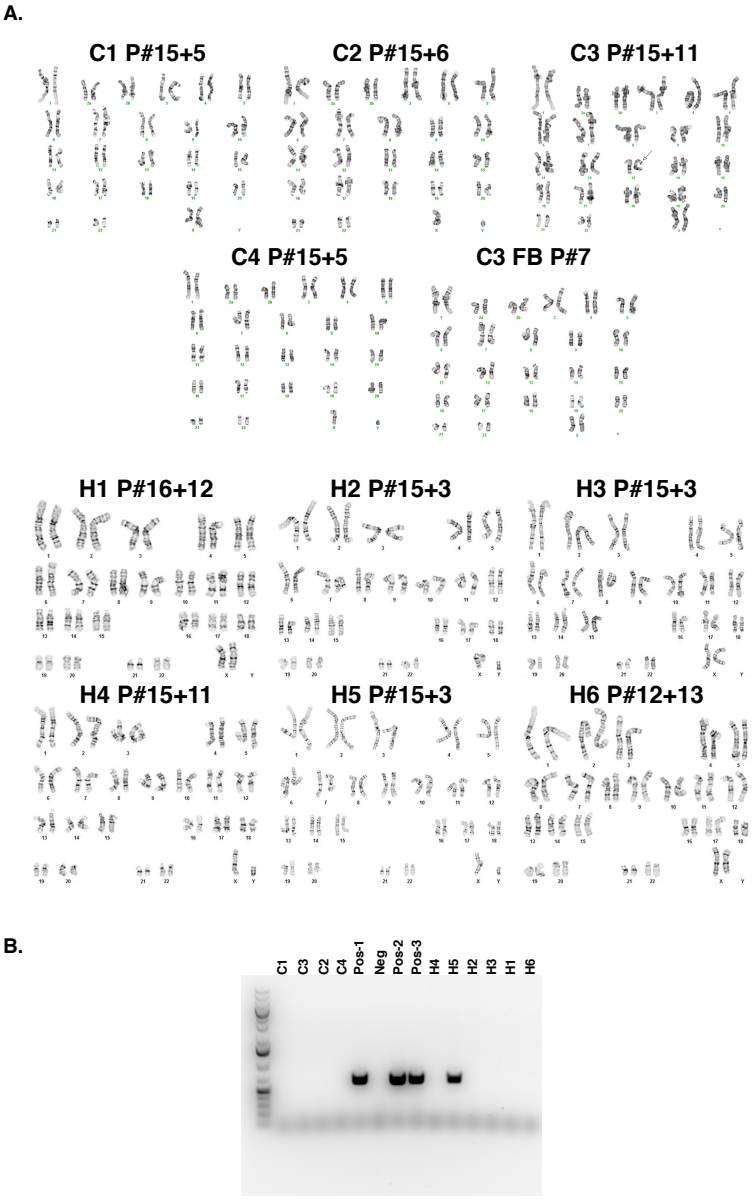


Figure 3.27: **Quality control for the iPSC lines used in this Study.** A. Karyotypes for human and chimpanzee iPSC lines. We identified additional bands in the p-arms of one chromosome 13 homolog and one chromosome 18 homolog for chimpanzee iPSC line C3. Thus we tested the source fibroblast line (C3 FB) to determine that these polymorphisms were normal polymorphisms and not formed de novo as a result of reprogramming. B. PCR gel to test for exogenous episomal reprogramming vectors in all iPSC lines used for this study. Pos indicates positive controls,Neg is a negative control. Human line H5 demonstrates a clear positive result for reprogramming plasmid.

3.5.2 Supplementary Tables

Table 3.1: **Table S1: RNA Integrity Number (RIN) for each of the 64 samples.** Column 1 contains the Sample ID. Column 2 contains the RIN score. The final row contains the outlier sample. Samples without RIN scores are annotated with an NA designation.

Table 3.2: **Biological and technical factors of interest with accompanying values for each sample.** This file contains sample information (including identifiers and demographics), experimental details, nucleic acid extraction information, library preparation information, sequencing information, and quality metrics. Row 1 contains the type of information. Row 2 contains a description of the variable, including units of measurement, if applicable. Row 3 lists the variable name. All other rows provide the values for each variable for the 63 non-outlier samples. Row 67 contains information for the outlier sample. B. A glossary of the technical factors and their definitions for this study.

Table 3.3: **TMM- and cyclic loess-normalized log2 counts per million for the 10,304 genes analyzed in this study for the 63 samples used in the downstream analysis.** This file contains the gene expression log2(CPM) values for the 10,304 genes after filtering lowly expressed genes and performing TMM- and cyclic loess normalization. Each row is an Ensembl gene name and each column is a Sample ID.

Table 3.4: **Purity estimates for each of the 64 samples based on gating.** rather than k-means clustering, both described in Methods. Column 1 contains the Sample ID. Column 2 contains the purity estimate. The final row contains the outlier sample. Samples without purity estimates were assigned an NA designation.

Table 3.5: **Results of our technical factor analysis** Column 1 contains the response variable and all other columns are explanatory variables in individual linear models. Unadjusted P values (A) and Benjamini-Hochberg adjusted P values (B) from regression models with PCs 1-5 as response variables and recorded biological and technical variables as explanatory variables (performed using 63 samples). Unadjusted P values (C) and Benjamini-Hochberg adjusted P values (D) from regression models with day and species as the response variables and technical factors of interest as the explanatory variables (63 samples). E. Benjamini-Hochberg adjusted P values from from regression models with PCs 1-5 as response variables and recorded biological and technical variables as explanatory variables. This analysis contained 40 data pointsthe data averaged over technical replicates– instead of the 63 individual samples.

Table 3.6: **Interspecies differential expression statistics from limma.** This includes the ENSG gene ID (Gene), Gene name, log2 fold change (logFC), average expression level (AveExpr), t-statistic (t), P value (P.value), q-value (adj.P.Val), and log-odds (B). Lists are in order of smallest to largest adjusted P value. A. Interspecies differential expression statistics from limma for day 0. B. Statistics for day 1. C. Statistics for day 2. D. Statistics for day 3.

Table 3.7: **Differential expression statistics from limma between days for each species.** A. Differential expression statistics from humans, day 0 to 1. B. Statistics from humans, day 1 to 2. C. Statistics from humans, day 2 to 3. D. Statistics from chimpanzees, day 0 to 1. E. Statistics from chimpanzees, day 1 to 2. F. Statistics from chimpanzees, day 2 to 3.

Table 3.8: **Information about the sharing of DE genes across species for a given transition (e.g. day 0 to 1).** Column 1 lists the days which were tested for DE genes in each species (Table S7). The pval refers to the Bonferroni-Hochberg adjusted P value from Table S7.

Table 3.9: **Differential expression statistics from limma testing a species-by-day interaction term.** A. Differential expression statistics for a human-by-day 1 interaction term. B. Differential expression statistics for a human-by-day 2 interaction term. C. Differential expression statistics for a human-by-day 3 interaction term.

Table 3.10: **Correlation motif assignments and enriched Gene Ontology (GO) annotations for different motifs.** A. Correlation motif assignments from Figure 5. The table of 8,004 genes and their correlation motif assignment (from Figure 5). 2,300 genes were not assigned to one of the eight correlation motifs and therefore are not included in this file. B. Enriched GO annotations in motifs 4 and 7. C. Enriched GO annotations in motif 2. D. Enriched GO annotations in motifs 3 and 5. E. Enriched GO annotations in each of the motif groups conditioned on genes present in the other motif groups.

Table 3.11: **The number of DE genes between species under various conditions.** Column 1 lists the number of samples and FDR cutoff. All tests used TMM- and cyclic loess-normalized log2(CPM) unless otherwise stated. Columns 2-10 list the total number of DE genes and day-specific DE genes at each day. The same voom+limma pipeline was used to generate all DE statistics (see Methods).

Table 3.12: **P values from model comparisons for each recorded technical variable.** Column 1 lists the technical variables. The technical factor type in Column 2 corresponds to the data type (1 = categorical variable with more than two levels, 2 = categorical variable with two levels, 3 = numerical variable), which impacted the model type (see Methods). P values were generated for models of the technical factors from days 0-1 (Column 3) and from days 0-3 (Column 4). Non-significant P values in all of the day 0-1 tests and most of the day 0-3 tests suggests that variation was relatively well distributed between the days and between the species.

Table 3.13: **A-B. Distributions of P values from F tests.** P value distributions from F tests of the null that variation in gene expression has not changed as samples progress throughout the time course. The alternative hypothesis in (A) is that there is a reduced variation in gene expression in the subsequent day (e.g. reduced variation at day 1 compared to day 0), whereas the alterative hypothesis is (B) is that there is an increase in variation in gene expression. C-E. Functional enrichments and phenotypes in our gene sets. C. Enriched GO annotations in genes with a shared reduction of variation in gene expression at the primitive streak relative to genes that do not show a reduction at this timepoint in either species. Mouse orthologs of the primate genes which we found a shared (D) reduction of variation in gene expression at the primitive streak or no reduction (E), and their associated phenotypes in null mutants.

Table 3.14: **100,000 estimated $\hat{\pi_0}$ values calculated from permuted P values obtained from F tests for reduced variation in gene expression levels.** Column 1 contains $\hat{\pi_0}$ estimates for the chimpanzee samples (based on permutations; mean = 0.274, standard deviation = 0.046) when one only considers genes for which reduced variation has already been in humans. Column 2 is the same data type but was calculated using human samples (based on permutations; mean = 0.825, standard deviation = 0.050) when considering only genes for which reduced variation has already been observed in the chimpanzees.

Table 3.15: **Log-likelihood values from Cormotif models.** Column 1 contains the seed number and Columns 2 and 3 are the log-likelihoods using 7 and 8 correlation motifs, respectively.

CHAPTER 4

A COMPARATIVE ASSESSMENT OF IPSC DERIVED CARDIOMYOCYTES WITH HEART TISSUES IN HUMANS AND CHIMPANZEES

4.1 Abstract

Comparative genomic studies in primates have the potential to reveal the genetic and mechanistic basis for human specific traits. These studies may also help us better understand inter-species phenotypic differences that are clinically relevant. Unfortunately, the obvious limitation on sample collection and experimentation in humans and non-human apes severely restrict our ability to perform dynamic comparative studies in primates. Induced pluripotent stem cells (iPSCs), and their corresponding differentiated cells, may provide a suitable alternative system for dynamic comparative studies. Yet, to effectively use iPSCs and differentiated cells for comparative studies one must characterize the extent to which these systems faithfully represent biological processes in primary tissues. To do so, we compared gene expression data from primary adult heart tissue and iPSC derived cardiomyocytes from multiple human and chimpanzee individuals. We determined that gene expression in cultured cardiomyocytes appear more similar to that of adult hearts than to other adult tissues, and we identified the optimal parameters for differentiation that minimized differences between cardiomyocytes and primary hearts. Using a comparative framework, we found that more than 50% of gene regulatory differences between human and chimpanzee hearts are also observed in cultured cardiomyocytes; conversely, inter-species regulatory differences seen in cardiomyocytes are found significantly more often in hearts than in other primary tissues. Our work provides a detailed description of the utility and limitation of differentiated cardiomyocyte as a system for comparative studies in primates.

4.2 Introduction

Comparative studies in humans and non-human apes are extremely restricted because we only have access to a few types of cell lines and to a limited collection of frozen tissues. In order to gain true insight into regulatory processes that underlie inter-species variation in complex phenotypes, we must have access to faithful model systems for a wide range of tissues and cell types. To address this challenge, we have previously established a panel of iPSCs from human and chimpanzee fibroblasts [42]. We can use these comparative iPSC panels to derive multiple cell types representative of the three main germ layers. For example, our lab recently differentiated the human and chimpanzee iPSCs into definitive endoderm cells [12]. Based on model organism studies of early development, we expected, and indeed confirmed, gene expression trajectories to be highly conserved during differentiation. Our hope is that employing iPSC based models from humans and chimpanzees will provide researchers with a dynamic and flexible system for comparative functional genomic studies using a large number of cell types. Towards this goal, the purpose of the current study is to evaluate the performance of one type of differentiated cell, cardiomyocytes, in a comparative context. This effort is not unique; there are quite a few previous studies that focused on characterizing similarities and differences between derived cell types to fetal and adult human and mouse tissues [8, 45, 64, 103, 118, 125, 145, 146, 147, 163]. Generally, results from these studies have demonstrated that the differentiated cultured cell types are most equivalent to fetal tissues rather than to the corresponding adult tissues. A few studies specifically explored protocol properties that may result in more mature derived cells [39, 69, 93, 95, 124, 145, 169, 172]. That said, the published works do not specifically address the properties that we set out to address with respect to the utility of differentiated cardiomyocytes in the context of a comparative study in primates. First, nearly all published studies were conducted using relatively few individuals (three or fewer), such that the observation of high similarity of gene expression patterns between cultured cells and primary tissue may be explained by

lack of statistical power. Second, published studies did not consider their observations in the broader context of other tissues or other species, so it is challenging to benchmark the observation of what is claimed to be small or large regulatory differences between tissues and cultured cells. Different protocols and batch effects make it difficult to perform meta-analysis of existing data to effectively address this issue. Finally, to date, no study that focused on the fidelity of differentiated cells included samples from chimpanzees. To address these gaps, we performed a comparative study that was specifically designed to allow us to effectively compare gene expression data between cultured cardiomyocytes to primary hearts from humans and chimpanzees. Our study was also designed to allow us to benchmark the results against regulatory differences seen between other primary tissues and across species. A key finding from previous studies pertaining to the fidelity of iPSC derived cell types is the importance of cellular maturation after terminal differentiation. While the initial steps of cardiomyocyte differentiation are fairly well developed, there remains some debate in the field as to the best method of cellular maturation [60, 148]. Common strategies include: treatment with triiodothyronine (T3) [52, 63, 168], electrical stimulation[93] and temporal maturation in culture. Maturation is likely to increase the fidelity of cardiomyocytes to primary adult heart tissue. Thus, to test the effects of temporal maturation and treatment with T3, we compared gene expression profiles from human and chimpanzee iPSC-derived cardiomyocytes at day 15, day 27 (with and without T3 treatment) with data generated from primary heart tissue from both species.

4.3 Results

The primary goal of this work is not to test a particular hypothesis, but to develop an understanding of the degree to which iPSC derived cardiomyocytes can serve as a model system for comparative functional genomic studies in primates. To do so, we used matched panels of 9 human and 10 chimpanzee iPSC lines; a subset of these iPSCs was previously described

[30, 42]. Standard functional characterizations of iPSC lines, quality controls, and metadata for the entire panel of samples are provided in Fig. 4.9-Fig. 4.11 and 4.1. We differentiated the iPSC lines into cardiomyocytes using a recently developed protocol [21, 22]. The same protocol was used for both the human and the chimpanzee iPSC lines. All of the cardiomyocyte cultures have started beating between the 6^{th} and the 8^{th} day. On days 20-27, a subset of cultures was treated with T3 (methods), which was shown to aid in the maturation of iPSC-derived cardiomyocytes [63, 69, 168]. We assessed the purity of the iPSC-derived cardiomyocytes using flow cytometry (Methods) for cardiac troponin I (TNNI) and cardiac troponin T (TNNT2). By day 15, the purity of our samples was on par relative to the standard in the field (average purity of 68%; 4.1). Importantly, we found no significant difference in sample purity across species (Fig. 4.12). We performed the cardiomyocyte differentiations in two batches, with partial sample overlap across the batches; thus, we performed technical differentiating replicates for a subset of samples. From the samples that were differentiated in the first batch, we collected RNA on day 27 from samples that were treated with T3. From the samples that were differentiated in the second batch, we collected RNA from the iPSCs, from differentiated cardiomyocytes on day 15, as well as on day 27, from samples that were either treated or not treated with T3. To benchmark the gene expression data that we collected from the iPSC and iPSC-derived cardiomyocytes against that of primary tissue, we also collected RNA from flash frozen heart tissues samples from 21 chimpanzees and 11 humans (Methods). Our overall study design (Fig. 4.12) may seem arbitrary and overall unbalanced, but it affords us the flexibility to choose to analyze different subsets of samples that are balanced with respect to different technical and biological factors. We can thus effectively compare within and between species gene expression primary from primary tissue and iPSC-derived cardiomyocytes that were harvested on days 15 and 27, with or without the treatment with T3. We examined and recorded the quality of the RNA samples from both tissues and cultures (4.1), and sequenced each sample to an average coverage of 49.4 million

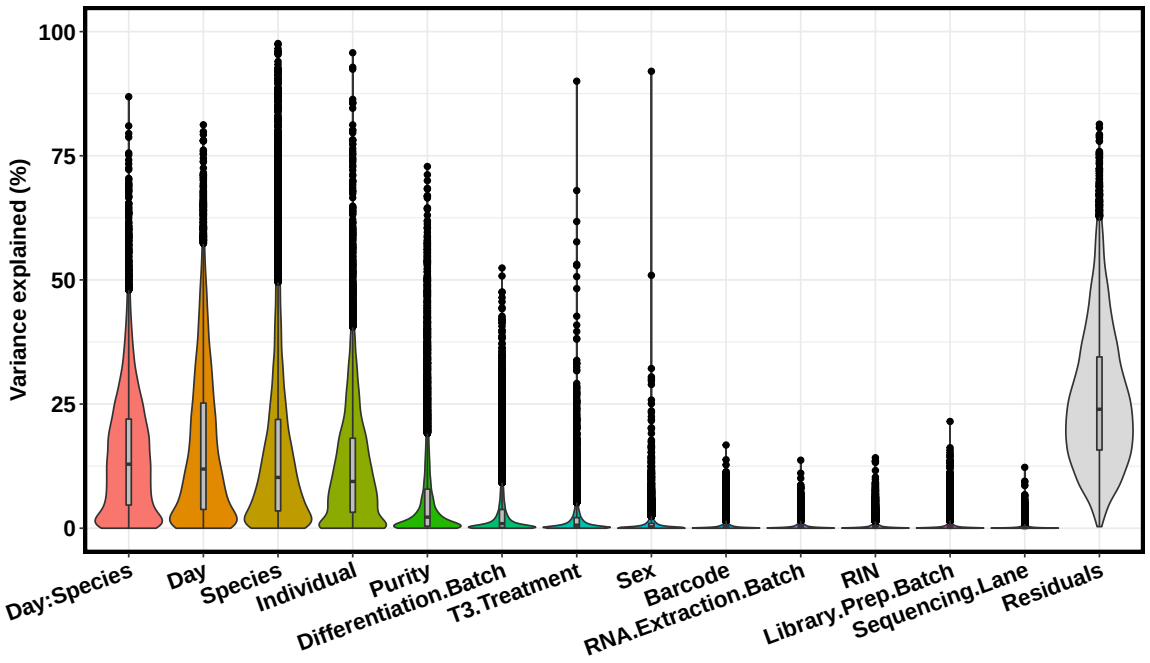


Figure 4.1: **Variance partition.** Results from variance partitioning analysis showing major drivers of gene expression within iPSC-CMs. All terms were modeled as fixed effects with the exception of individual which was modeled as a random effect.

total reads. We mapped the sequencing reads to an updated two-way human-chimpanzee orthologous exon reference set (Methods) and estimated relative gene expression levels using edgeR [115]. Throughout the differentiation and sample processing steps, we have recorded a large number of biological and technical properties (sample metadata available in 4.1). As a first step of our analysis, we determined the potential impact of different properties of the study design and sample metadata on the observed gene expression levels (using variance partition; see Methods). Most study design properties explain only a modest amount of variation, but as expected, technical batch is associated with substantial amount of variation, as is sample purity (Fig. 4.1). However, neither sample purity nor technical batch are associated with species (Fig. 4.13).

4.3.1 Which differentiated cardiomyocytes resemble primary hearts?

Overall, we collected data from 110 samples (19 iPSCs, 59 differentiated cardiomyocytes, and 32 primary heart samples). To examine global trends in gene expression levels, we normalized the combined data across all samples (using TMM; Methods), and reduced dimensionality for visualization using principal components analysis (PCA; Fig. 4.2A). We found that the major source of variation in the data is cell type; PC1 and PC2, which explain 37% and 21% of the variance, respectively, are highly associated with cell type ($P < 10^{-16}$). The next major source of variation, as expected, is species (which is strongly associated with PC3, which explains 7% of the variance; $P < 10^{-16}$). We performed unsupervised hierarchical agglomerative clustering on the correlation matrix of gene expression data, and found that the major trends seen in the PCA are robust with respect to clustering method (Fig. 4.2B). We noted that data from iPSC-derived cardiomyocytes clusters by species first, then by maturation day or treatment.

We focused on the differences between iPSC-derived cardiomyocytes at days 15 and 27, with and without the treatment with T3. For this analysis, we considered within-species gene expression patterns. Using the data from each species independently, we asked which cardiomyocyte cultures are more similar to adult heart tissue with respect to their gene expression profiles. We first focused on temporal maturation (the untreated samples at days 15 and 27) and found that average pairwise correlation in gene expression data is higher between untreated day 27 cardiomyocytes and hearts than between day 15 cardiomyocytes and hearts, regardless of species (Fig. 4.3A, $P < 10^{-16}$).

This result indicates that temporal maturation increases the similarity between cardiomyocytes and adult heart tissue for both species. We proceeded to compare gene expression data between heart tissues and cardiomyocytes cultures at day 27, which were treated or not treated with T3. We found that gene expression data from samples treated with T3 are more similar to that of primary heart tissues, regardless of species (Fig. 4.3A, $P < 10^{-4}$).

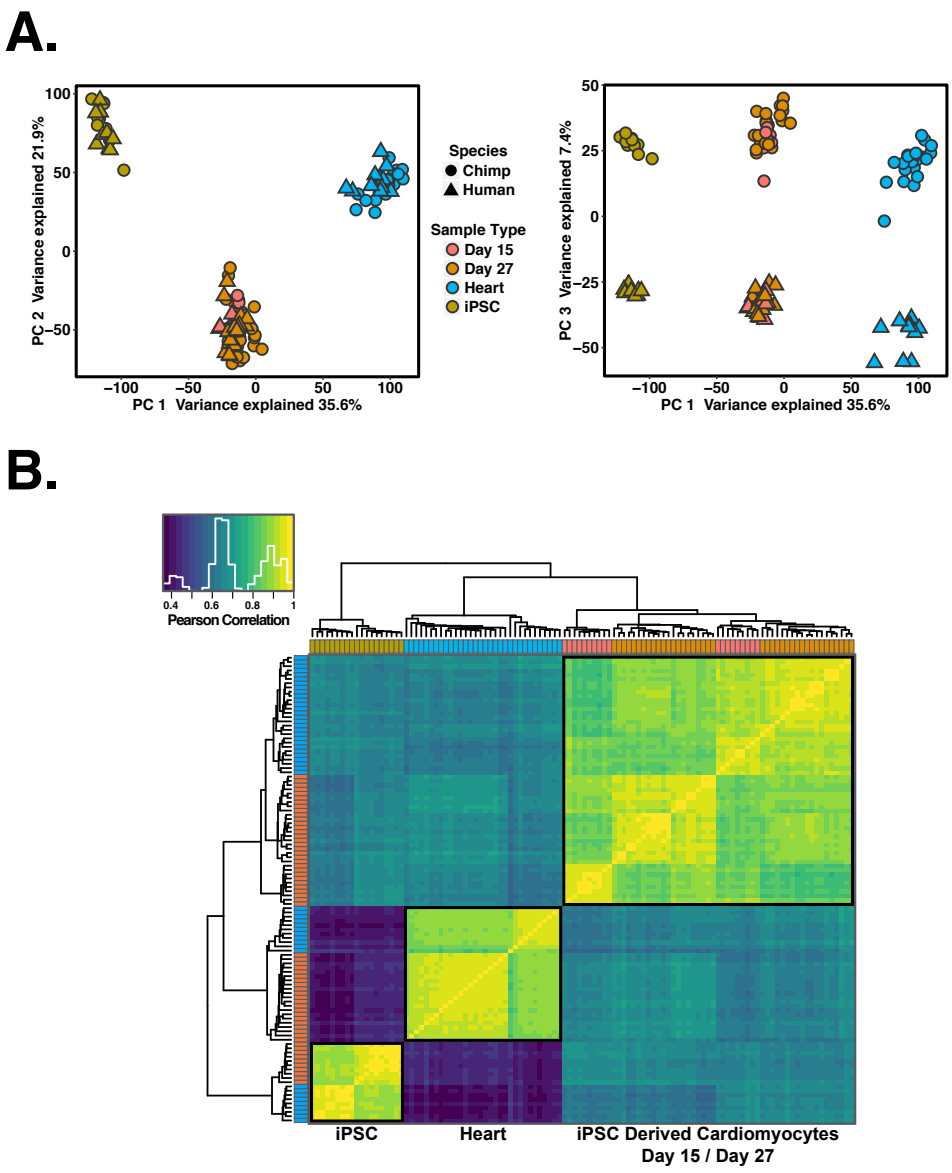


Figure 4.2: **Global patterns in gene expression across data collected in this study.** A. Normalized $\log_2$(RPKM) expression measurements for all genes projected onto the axes of the first two principal components. Color indicates cell or tissue type. Shape represents species. PC1 and PC2 are both strongly associated with cell or tissue type, PC3 is strongly associated with species. B. Heat map of the Pearsons correlation matrix of normalized $\log_2$(RPKM) gene expression values from 13,878 genes. Each square represents the Pearsons correlation of the normalized expression values between two samples.

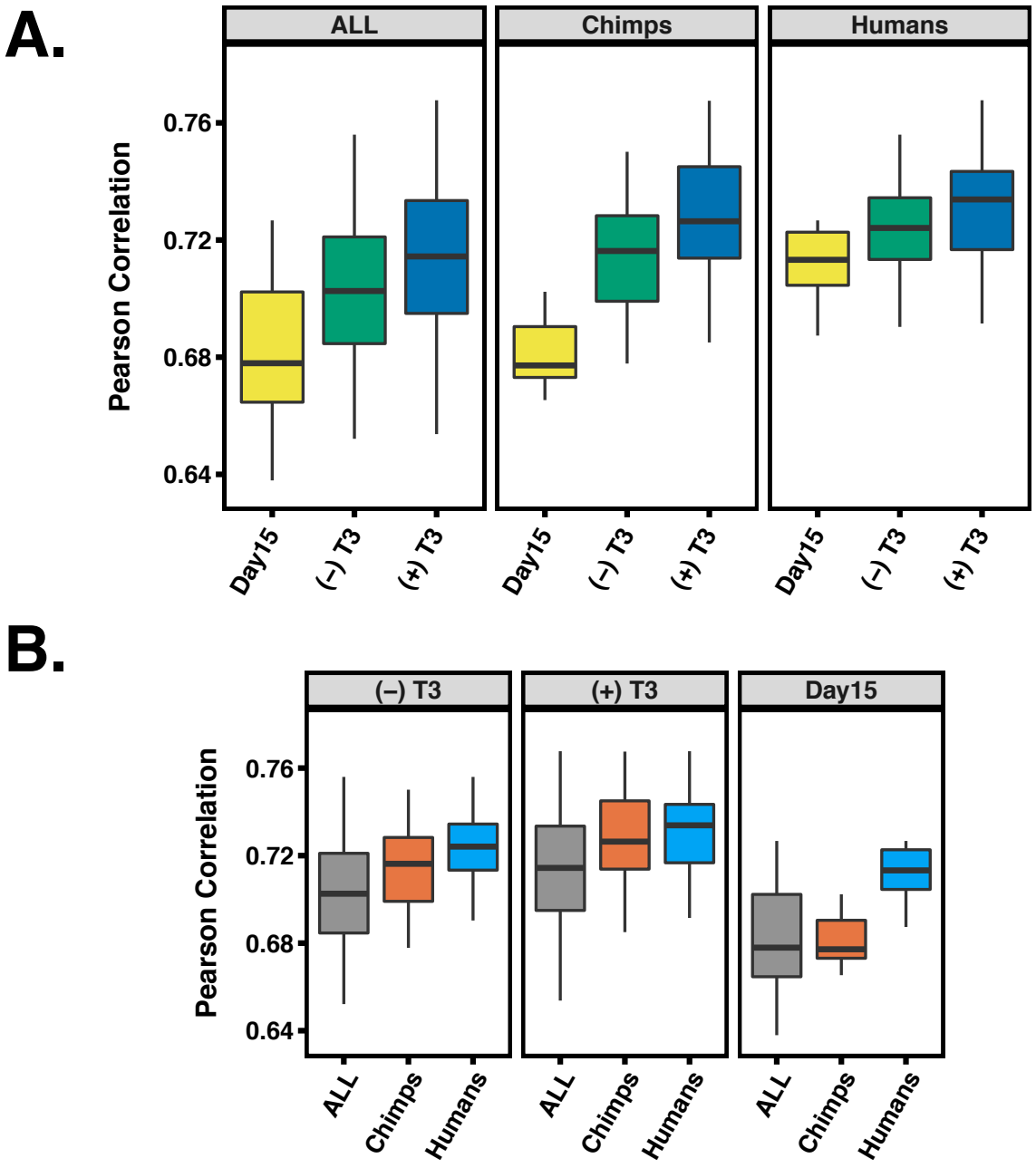


Figure 4.3: **Pairwise correlations between heart tissue and different iPSC derived cardiomyocytes.** A. Pearson correlations between heart tissue, day 15, (-)T3 and (+)T3 day 27 iPSC derived cardiomyocytes for chimpanzee, humans and All (combined, species not separated). B. Same data as in A, but reordered to show differences within a treatment across species.

We compared the maturation effects across species; namely, we asked whether hearts and cardiomyocytes are more similar to each other in humans or chimpanzees. We found that, at day 15, human cardiomyocytes are more similar to human adult heart tissue than chimpanzee cardiomyocytes are to chimpanzee adult heart tissue (Fig. 4.3B, $P < 10^{-7}$). However, by day 27, we do not observe any significant difference across species in this comparison ($P = 0.133$).

4.3.2 Which genes are differentially regulated between hearts and cultured cardiomyocytes?

Based on our observations we concluded that as the cells mature, the fidelity of iPSC-derived cardiomyocytes as a model for primary hearts increases in both species. Moreover, T3 treatment helps to mature iPSC derived cardiomyocytes, altering their gene expression such that it is more similar to that of adult tissues than temporal maturation alone. Yet, even T3 treated samples at day 27 are clearly distinct from heart samples (Fig. 4.2 and Fig. 4.3). We thus characterized the specific regulatory difference between hearts and iPSC derived cardiomyocytes. To do so, we considered a subset of the data, from iPSC-derived cardiomyocyte samples from 7 humans and 7 chimpanzees, all with purity above 50% (Fig. 4.14). We used a linear model framework to perform a combined analysis of all data from these samples, as well as from the corresponding iPSCs, and from the primary heart tissues (Methods). We focused on genes that were classified as differentially expressed between cell types in both humans and chimpanzees (Methods). At an FDR of 5% we classified 3,446 differentially expressed genes when we compared data from samples at day 27 and heart tissue, and 4,115 differentially expressed genes when we compared data from samples at day 15 and heart tissue. Using the same approach, we classified nearly 8,000 genes when we compared data between iPSCs and cardiomyocytes at either day 15 or day 27. To gain insight into the genes that are differentially regulated between primary hearts

and iPSC-derived cardiomyocytes, we considered GO functional annotations. We classified differentially expressed as those that are expressed at higher level in either primary hearts or the iPSC-derived cardiomyocytes, and further, to those that are practically not expressed in either primary hearts or the cardiomyocytes (based on our filtering criterion for lowly expressed genes that was used throughout the study; Log_2 CPM < 2; Methods). Among genes that are expressed in hearts but not in cardiomyocytes, we found an almost exclusive enrichment for biological processes relating to immune responses and inflammation. In turn, among genes that are expressed in cardiomyocytes but not in hearts, we found enrichment of genes involved in cell cycle regulation. Among genes expressed in both cardiomyocytes and hearts, we found that genes involved in transcription regulation tend to be expressed at higher levels in cardiomyocytes and genes involved in metabolic processes specifically related to lipid oxidation tend to be expressed at higher levels in hearts (Fig. 4.4; complete enrichment results in 4.3; all reported enrichments associated with FDR < 0.05). Remarkably, when we consider all of these enrichment results, we are able to account for 67% of genes that are classified as differentially expressed between hearts and iPSC-derived cardiomyocytes.

Up to this point, we considered gene expression patterns within species. To specifically examine iPSC-derived cardiomyocytes as a model system for comparative genomic studies, we proceeded to compare inter-species regulatory differences in either cell cultures or their corresponding primary tissues. We used a framework of linear modeling as above, but this time focused on the species by cell type interaction effect. To account for incomplete power when we classified genes as differentially expressed between species in any given cell type, we used a relaxed statistical cutoff for secondary observations. Namely, when genes were classified as differentially expressed between species in one cell type using a stringent statistical cutoff (e.g., FDR of 5%), we used a relaxed statistical cutoff to classify the same gene as differentially expressed between species in another cell type (e.g., nominal $P < 0.05$; see Methods). Using this approach, we found that roughly 50% of genes that are

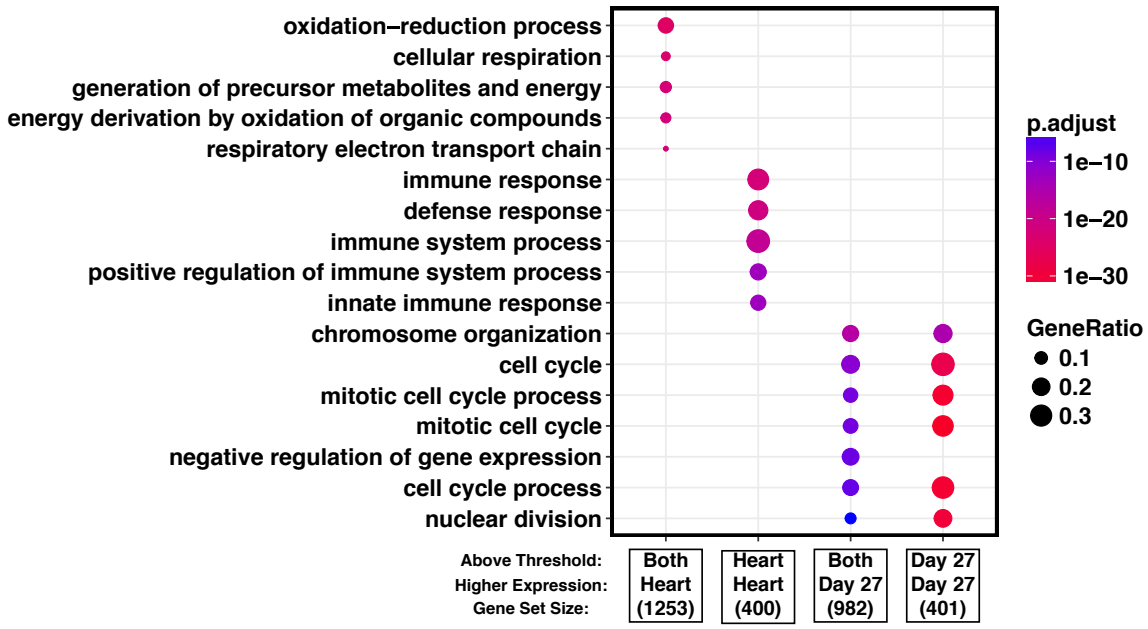


Figure 4.4: **Gene enrichment results for genes that are DE between day 27 cardiomyocytes and heart tissues.** Genes identified as DE between day 27 cardiomyocytes and heart tissues were split into 4 groups based on whether they were above the gene filtering threshold ($\log_2$CPM > 2) in both day 27 and heart tissues (Both) or below in one or the other (Heart indicates that the gene is above the threshold in heart tissue, Day 27 indicates the gene is above threshold in Day 27 iPSC-CMs) or in day 27 respectively). Each gene set was further subset into where a DE gene had higher expression, higher in heart tissue (Heart) or higher in day 27 (Day27). Enrichment in GO biological processes for each of the lists is shown, Gene list sizes are shown under the label for each gene set.

classified as differentially expressed between human and chimpanzee in heart tissues are also differentially expressed between cultured cardiomyocytes of the two species, regardless of whether we considered cardiomyocytes that were harvested at day 15 or day 27 (Fig. 4.5). This result is robust with respect to the degree of purity of the cardiomyocyte cultures (within a range of 40-70%; Fig. 4.15). When we considered genes that are classified as differentially expressed between the species either in hearts or in cardiomyocytes (Methods), we found enrichment in similar terms for both cell types (4.4). Specifically, we found enrichment in terms relating to regulation of metabolic processes, RNA metabolic processes, and regulation of gene expression. When we considered genes that are classified as differentially expressed in both hearts and in cardiomyocytes, we found an enrichment of terms relating to animal organ development, cellular developmental processes and positive regulation of transcription (4.4).

4.3.3 Cardiomyocytes are more similar to hearts than to other primary tissues

Our observations thus far indicate that while much of the heart regulatory divergence between species can be captured by using iPSC-derived cardiomyocytes, there are also thousands of genes whose expression differs between hearts and cultured cardiomyocytes. To provide context for these observations, we used data from 15 publicly available RNA-seq studies, which include samples from four different adult tissues from both humans and chimpanzees, as well as from human fetal tissues [7, 18, 46, 70, 76, 102, 107, 120, 122, 144, 154, 158, 159, 167]. We processed all the public data and our own newly collected data (from a total of 231 human samples and 111 chimpanzee samples) using a common analysis pipeline and modeled the data together (see Methods).

We first focused on similarities and differences between tissues and cell types. We performed PCA using data from all 342 samples, which visually suggested separation of the data

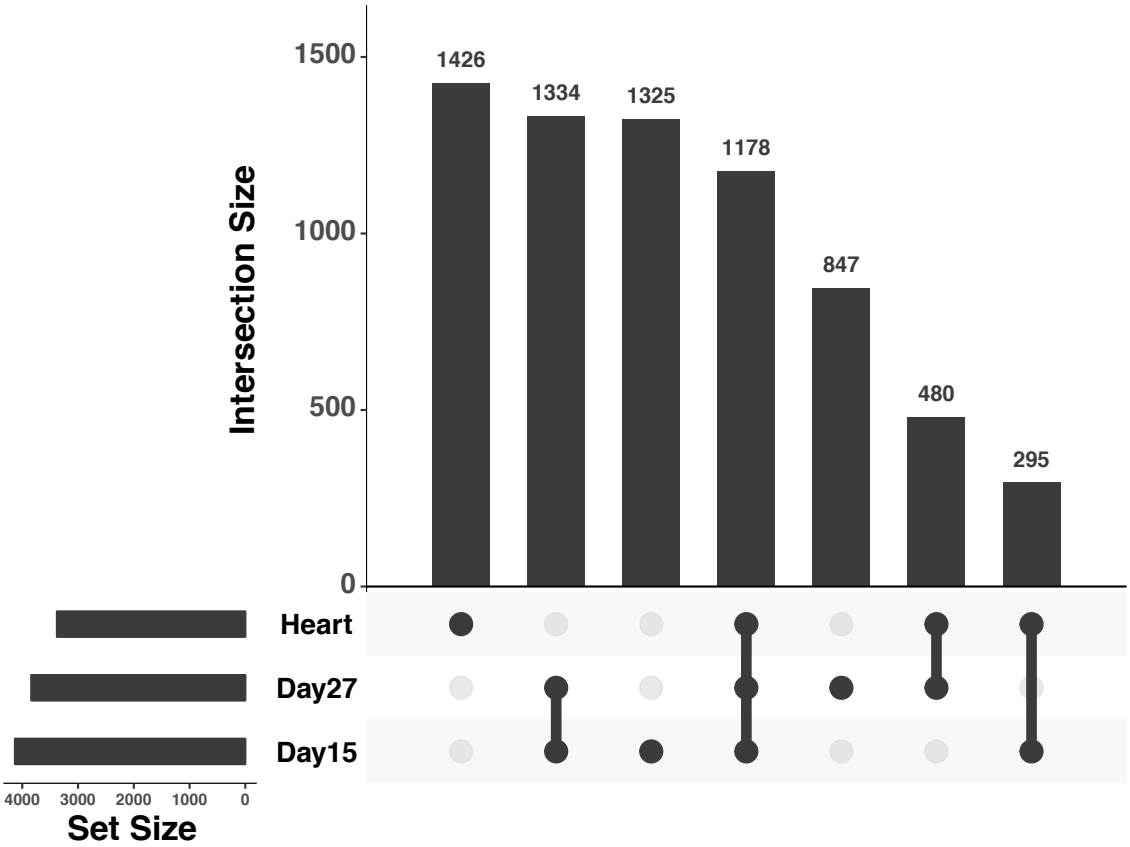


Figure 4.5: **Upsetr plot showing set overlaps across multiple sample types for genes identified as DE between Human and Chimp.** Total set sizes are shown on the bottom left, overlaps are shown by links with a filled circle, bar plot above link shows size of a specific overlap. Plot generated using R package Upsetr ([33]).

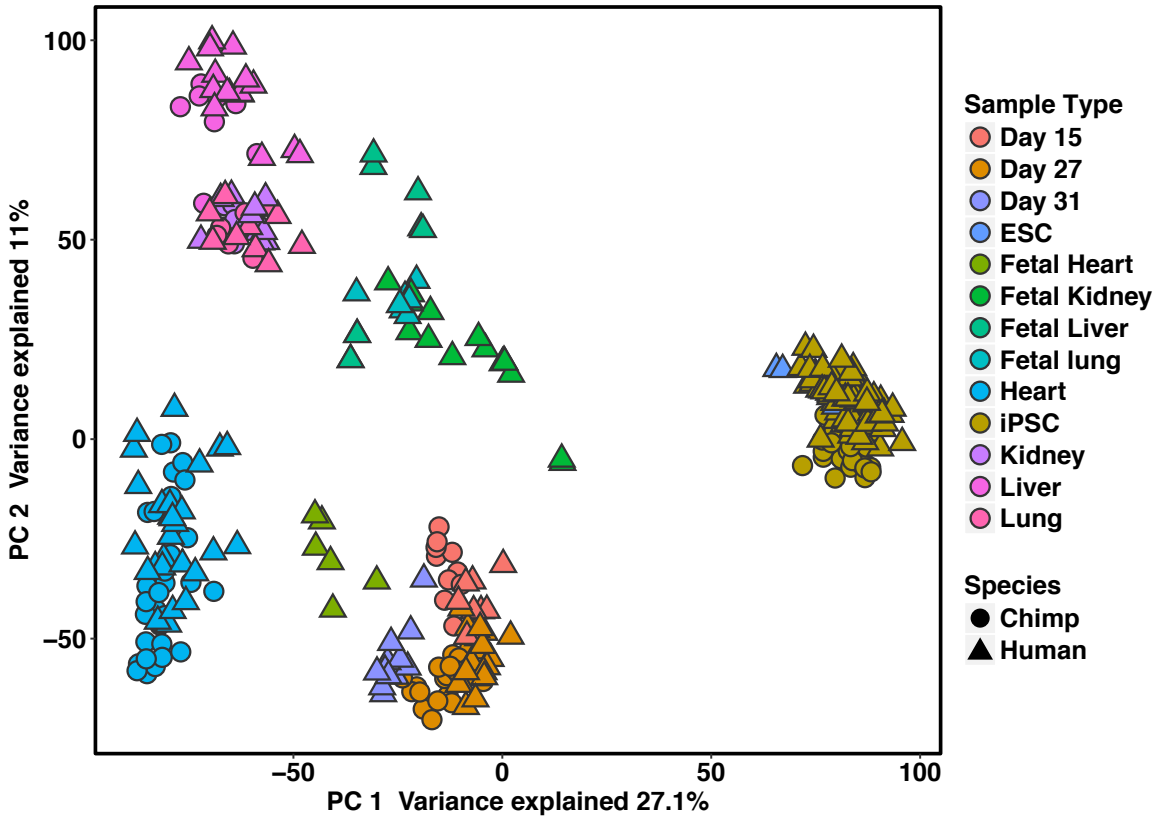


Figure 4.6: **Global patterns using data from multiple studies from adult and fetal tissues.**Normalized $\log_2$(RPKM) expression measurements for all genes projected onto the axes of the first two principal components. Color indicates cell or tissue type. Shape represents species. PC1 and PC2 are both strongly associated with cell or tissue type.

by cell and tissue type in the first several principal components (Fig. 4.6; PC1 and PC2 are both highly associated with cell type or tissue type; $P < 10^{-16}$). We only observed an association with species when we considered the loading on PC6 ($P < 10^{-16}$). We performed unsupervised hierarchical agglomerative clustering using the correlation matrix of the gene expression data across samples (Fig. 4.7). The data clusters by cell and tissue type followed by species (with the exception of public data from one chimpanzee heart sample from [102]). Notably, we found that data from iPSC-derived cardiomyocytes are consistently more similar to data from fetal heart tissue, followed by data from adult heart tissues, than any other non-heart tissues. These observations are consistent with the notion that iPSC-derived cardiomyocytes are most similar to first trimester fetal hearts [147].

We next evaluated the ability to iPSC-derived cardiomyocytes to capture regulatory divergence in hearts in the context of comparative data sets from other tissues. To do so, we used RNA-seq data collected from four individuals from each species from four different tissues: heart, lung, liver and kidney (Methods). We classified genes as differentially expressed between human and chimpanzee in each tissue and in day 27 cardiomyocytes using identical analysis pipelines (Methods). To account for incomplete power when we classified genes as differentially expressed between tissues and cell types, we allowed for a more relaxed statistical cutoff for secondary observations (Methods). As expected, a large number of genes show inter-species differences in expression level across all tissues while different subsets of genes are differentially expressed between the species in only a single tissue (Fig. 4.16). There are as many genes that are differentially expressed between species exclusively in differentiated cardiomyocytes as in all the other primary tissues. This may reflect true biological patterns or the artificial nature of the cultured cardiomyocytes - our study is not designed effectively to effectively address this question. Importantly, genes that are differentially expressed between human and chimpanzee in cardiomyocytes are much more likely to also be classified as differentially expressed between the species in hearts than in any other tissue (Fig. 4.8).

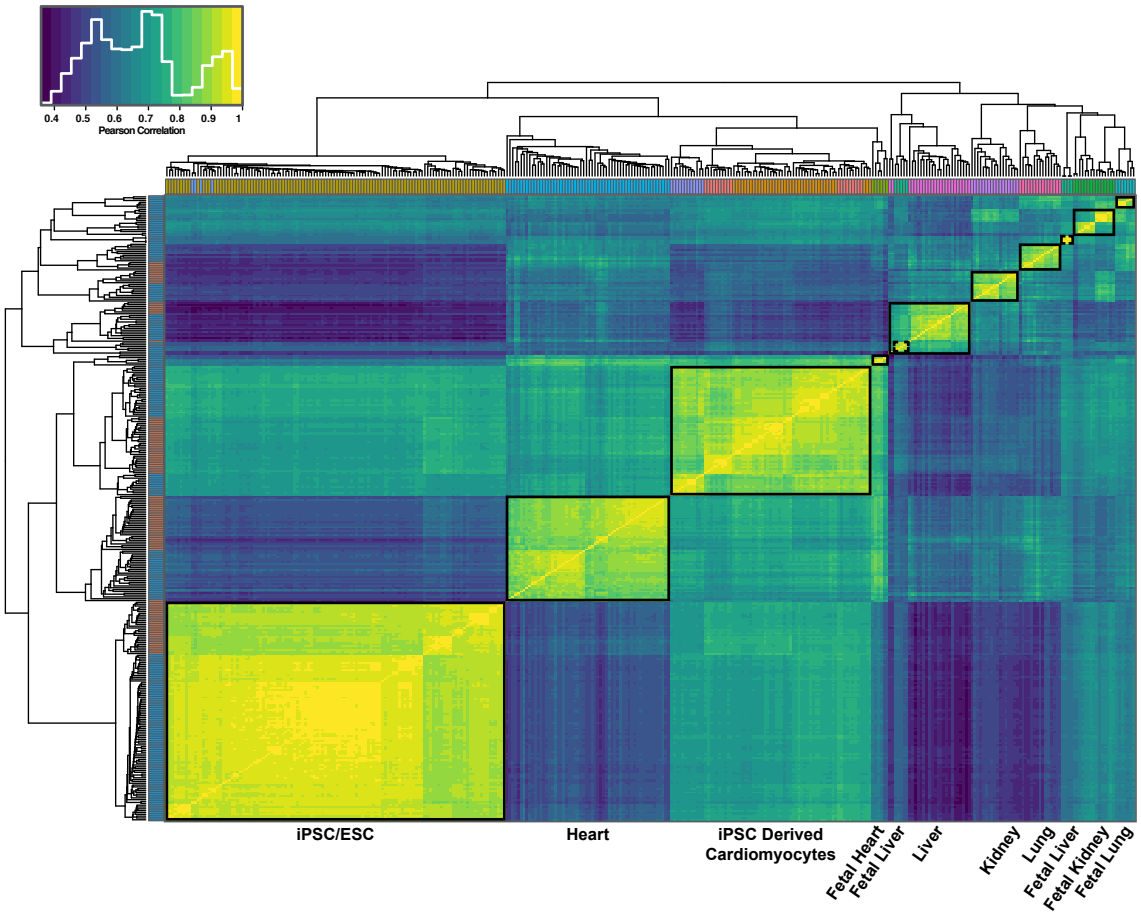


Figure 4.7: **Clustering data from multiple studies from adult and fetal tissues.** Heat map of the Pearsons correlation matrix of normalized $\log_2$(RPKM) gene expression values from 335 samples for 15,829 genes. Each square represents the Pearsons correlation of the normalized expression values between two samples.

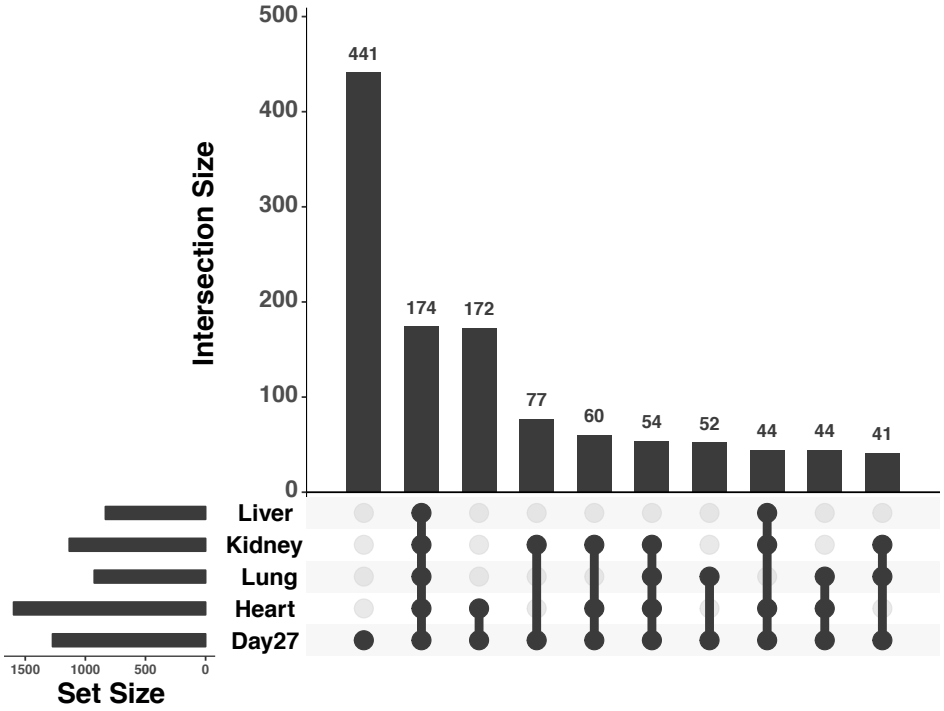


Figure 4.8: **Upsetr diagram showing benchmarking results for iPSC derived cardiomyocytes using an independent tissue reference set.** Total set sizes are shown on the bottom left, overlaps are shown by links with a filled circle, bar plot above link shows size of a specific overlap. Plot generated using R package Upsetr ([33]).

4.3.4 Discussion

The development of iPSC based model systems have the potential to transform the field of comparative primate genomics, assuming that differentiated cells can recapitulate biological processes that occur in primary cell types. Quite a few studies focused on gene regulatory variation in *in vitro* differentiated cells have already been published (including by our group), but a systematic assessment of the fidelity of these differentiated cells with respect to primary tissue has not been performed. The current study was therefore designed to specifically assess the degree to which differentiated cardiomyocytes can serve as a model system in which to study regulatory differences in human and chimpanzee hearts. We debated which cell type to use for this first exploration of the fidelity of differentiated cells. Many different cell types are accessible using iPSC differentiation protocols [22, 31, 73, 88]. Despite the seemingly many options, in most cell types, individual variation still remains a current barrier to generate consistently pure cultures of target cell types [21, 55, 61, 89, 99, 156]. In contrast, cardiomyocyte differentiation protocols are relatively robust with respect to inter-individual variation [21]. This outcome is partially due to the dominance of a single pathway (WNT signaling) in resolving developmental bifurcations in favor of cardiomyocyte generation in vitro [22, 110]. An additional advantage of utilizing a cardiac lineage for benchmarking is the reasonably homogenous nature of cardiac muscle (e.g., the composition of the left ventricle is upwards of 40% cardiomyocytes [173]). With these factors in mind, differentiations of iPSCs to cardiomyocyte provide an optimal framework to begin characterizing the similarity of iPSC derived cell types to primary tissue in primates. We did not expect differentiated cardiomyocytes to be identical to hearts. That would not be a realistic expectation. We hoped that differentiated cardiomyocytes would be most similar to hearts than to any other tissue, and we hoped to shed light on the opportunities and limitations associated with the use of differentiated cardiomyocytes to study hearts. Our observations confirmed that gene expression data from differentiated cardiomyocytes are more similar to data from hearts

than any other tissue we tested, but we also found thousands of differentially expressed genes between hearts and differentiated cardiomyocytes. The majority of these differentially expressed genes, however, may not point to inherit limitations of differentiated cardiomyocytes as a model, but rather they help us better understand the model properties. Indeed, 67% of differentially expressed genes between hearts and differentiated cardiomyocytes may be explained, not by artifacts specific to the model, but by expected biological differences between *in vitro* cultured relatively pure cells and a complex primary adult tissue. While most cells in the adult hearts no longer proliferate and do not have a regulated cell cycle, the differentiated cardiomyocytes, which resemble embryonic more than adult cells, are still likely to have not yet completely exit the cell cycle. It is thus may be expected that genes with functions related to transcription regulation and those involved in cell cycle regulation (there is a high overlap between these categories) would be more highly expressed in young differentiated cardiomyocyte cultures than in adult hearts. The cultured cells are relatively pure, while the hearts contain a complex composition of cells, including blood cells and resident immune cells. It is therefore reasonable that genes involved in immune responses and inflammation are highly expressed in hearts compared with differentiated cardiomyocytes. Finally, it has been previously reported that the metabolic state of hearts and differentiated cardiomyocytes are different [145]. Consistent with that report, we found that genes involved in lipid metabolism are highly expressed in hearts relative to differentiated cardiomyocytes, an exacted observation given that our cell cultures were not supplemented with lipids. Thus, the majority of the thousands of differentially expressed gene between primary hearts and cardiomyocytes are not necessarily indicative of inherit limitation of the *in vitro* model. Rather, many of these regulatory differences point to biological (cell composition) and environmental (lipids) differences between cell culture and primary tissue. When we specifically considered inter-species comparisons of gene expression levels using either cardiomyocytes or hearts, we found similar patterns. Namely, while most genes are classified as differen-

tially expressed between humans and chimpanzee in both differentiated cardiomyocytes and hearts, thousands of genes are differentially expressed between the species in either hearts or cardiomyocytes. Yet, the discordant inter-species patterns in cardiomyocytes and hearts mostly involve genes with functions related to metabolic processes, regulation of gene expression, and cell cycle. As we discussed above, there are likely biological and environmental explanations to these patterns. We thus conclude that iPSC derived cardiomyocytes are a useful model with which to conduct comparative studies of gene regulation. We expect that iPSC derived cardiomyocytes will be mainly used to comparatively study dynamic gene regulatory processes, such as during differentiation, when frozen heart samples even when available are unsuitable. The model is not perfect, but our study indicates that regulatory patterns observed in cultured cardiomyocytes cannot be mistaken as anything other than a representation of the biological processes that occur in primary hearts. With that in mind, we recommend that the data from our study, which are available in unprocessed and processed forms, should be used to evaluate the likelihood that future observations based on cultured cardiomyocytes faithfully represent regulatory patterns in hearts.

4.4 Methods

4.4.1 Human and chimpanzee iPSC panel

In this study, we include chimpanzee and human iPSC lines from previously described panels [42, 30, 12] matched for cell type of origin, reprogramming method, culture conditions. Characterization for all lines used is provided in Fig. 4.9-Fig. 4.11. Original chimpanzee fibroblast samples for generation of iPSC lines were obtained from the Yerkes Primate Center under protocol 00612. Human fibroblasts samples for generation of iPSC lines were collected under University of Chicago IRB protocol 110524. Feeder free iPSC 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 a 50/50% ratio of iDEAL/E8 feeder free medium that was prepared in house as specified previously [78]. Cell culture was conducted at 37C, 5% CO_2, and atmospheric O_2.

4.4.2 Cardiomyocyte differentiation

Cardiomyocytes were generated following a recently protocol ([21, 22, 109]) with minimal modification. At 12 hours prior to initiating differentiation, iPSC lines at 70-90% confluence were seeded to achieve a starting density of 150,000-250,000 cells/cm^2. Differentiations were initiated by removing stem cell maintenance media and adding RPMI base media (with HEPES LifeTechnologies# 22400105) supplemented with 6μM CHIR9902 (LClabs# C-6556), 0.5 mg/mL BSA (Sigma-Aldrich# A2153), 0.213 mg/mL ascorbic acid (Santa Cruz Bio# sc-228390), 0.3μg/mL sodium selenite (Sigma-Aldrich# S5261), 5μg/mL Holo-transferrin (Sigma-Aldrich# T3705), 1μg/mL Linolenic acid (Sigma-Aldrich# L2376), 1μg/mL Linoleic acid (Sigma-Aldrich# 1012) and 1x Pen/Strep (Days 0-1 media). After 48 hours the media was changed to RPMI base media (with HEPES LifeTechnologies# 22400105) supplemented with 2μM Wnt-c59 (Tocris# 5148), 0.5 mg/mL BSA (Sigma-Aldrich# A2153), 0.213 mg/mL ascorbic acid (Santa Cruz Bio# sc-228390 and 1x Pen/Strep (Days 2-3 media). After another 48 hours, the media was changed to RPMI base (with HEPES LifeTechnologies# 22400105) supplemented 0.5 mg/mL BSA (Sigma-Aldrich# A2153), 0.213 mg/mL ascorbic acid (Santa Cruz Bio# sc-228390) and 1x Pen/Strep. Media was changed every 48 hours until day 10 (total of 2 additional media changes). At day 10 the media was switched to RPMI media (no glucose, Cellgro#10-043-CV) supplemented with 5 mM Sodium D/L Lactate (Sigma-Aldrich# L4263), 0.5 mg/mL BSA (Sigma-Aldrich# A2153), 0.213 mg/mL ascorbic acid (Santa Cruz Bio# sc-228390) and 1x Pen/Strep, media was changed every 48 hours (total of 2 additional media changes). At day 15, cells were dissociated using TrypLE and split to a density of 0.4 x 10^6 cells/cm^2 onto Matrigel coated dishes. Cells were replated in RPMI

media (no glucose, Cellgro#10-043-CV) supplemented with 5% FBS, 10 mM D-galactose (Sigma-Aldrich# G5388), 1 mM Sodium Pyruvate (LifeTechnologies# 11360070), 1X NEAA (LifeTechnologies# 11140076), 5 mM HEPES (LifeTechnologies# 15630080), mM Sodium D/L Lactate (Sigma-Aldrich# L4263), 0.5 mg/mL BSA (Sigma-Aldrich# A2153), 0.213 mg/mL ascorbic acid (Santa Cruz Bio# sc-228390) and 1x Pen/Strep (Galactose media). After 24 hours, the media was changed to Galactose media without FBS. Media was changed every 48 hours (total of 1 additional media change). At day 20 media was changed to Galactose media supplemented with 3 ng/mL Triiodothyronine (T3, Sigma-Aldrich# T6397). Media was changed every 48 hours until cells were harvested at day 27 (total of 3 additional media changes). For the first two media changes (Days 0 and 2) cultures were supplies with an excess of media (0.499 mL/cm^2), all other days media was added at (0.2495 mL/cm^2).

4.4.3 Determining cardiomyocyte purity using flow cytometry

Day 15 cells were dissociated as described above, and aliquots were pulled off for flow analysis prior to replating. Day 27 iPSC-CMs, were dissociated by 10 minute incubations with TrypLE supplemented with 0.5 U/ml Liberase TH. Dissociated cells were centrifuged at 200 x g for 5 minutes at 4°C and washed with PBS. Cells were fixed and stained as specified previously ([21, 22]). Briefly, 0.5-1 million cells were fixed using a 1% PFA solution. Cells were fixed at 4°C for 30 minutes before washing once using FACS buffer (autoMACS Running Buffer, Miltenyi Biotech). Fixed cells were permeabilized by incubating with cold 90% methanol for 30 minutes at 4°C. Fixed, permeabilized cells were washed 2x using FACS buffer prior to staining. For immunostaining, 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 0.5% BSA/0.1% Triton X-100, PE-labeled TNNT2 (Bdbio clone 13-11) and Alexa Flour 647 labeled-TNNI3 (BDbio clone C5) All antibodies were used at the manufac-

turer recommended dilution. Cells were stained for 1 hour at 4°C and subsequently washed 3x with a PBS solution containing 0.5% BSA/0.1% Triton X-100. Stained cells were resuspended after the final spin in 150μL FACS buffer containing 5μM Vybrant DyeCycle Violet Stain (Thermofisher #V35003) for acquisition on a BD LSR II flow cytometer. After data acquisition, we used the program FlowJo to determine compensation scaling. To do so, we used 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. To determine gating for dual positive iPSC-CMs, we used a biological negative (an iPSC line stained in parallel) to determine appropriate positive signal for gating

4.4.4 Isolation of RNA and sequencing

To isolate RNA and DNA from cardiomyocytes and iPSC lines, we first aspirated old cell culture media, then rinsed culture wells using PBS. After aspirating PBS, 300μL of DNA/RNA lysis buffer was added directly to culture plates to lyse cells on the plate. The entire lysate was transferred to a 1 mL tube and immediately frozen for extraction with other samples in balanced batches. We extracted the RNA using the ZR-Duet DNA/RNA MiniPrep kit (Zymo) with the addition of an on column DNAse I treatment step prior to RNA elution. Post-mortem human heart tissues were provided by the National Disease Research Interchange (NDRI). Chimpanzee post-mortem heart tissues were provided by Yerkes primate center, the Southwest Foundation for Biomedical Research, and the New Iberia Research Center, MD Anderson Cancer Center. For lysis of heart tissue, frozen tissue from the left ventricle was broken down into smaller chunks while working on dry ice to prevent thawing, and subsequently weighed in 1.5 mL tubes to determine weight of tissue. For each mg of tissue, we added 8μL of 1X RNA/DNA shield buffer. Tissues were homogenized in RNA/DNA shield using sterile plastic pestles. To further homogenize tissues and release RNA and DNA, tissues were digested with Proteinase K for 30 minutes at 55°C. Digested, homogenized tis-

sues were centrifuged for 2 minutes to pellet debris, the volume of supernatant was measured and was transferred to fresh tubes. An equal volume of DNA/RNA lysis buffer was added and the solution was mixed prior to freezing at -80°C. Frozen samples were thawed and extracted in parallel using the same protocol as used for cells. We used non-strand specific, polyA capture to generate RNA-seq libraries according to the Illumina TruSeq protocol. To estimate the RNA concentration and quality, we used the Agilent 2100 Bioanalyzer. We added barcoded adaptors (Illumina TruSeq RNA Sample Preparation Kit v2) and sequenced the 100 base pair single-end RNA-seq libraries on the Illumina HiSeq 4000 at the Functional Genomics Core at University of Chicago on two flow cells.

4.4.5 Updating Orthologous Exons

As part of the work contained in this paper, we updated our previously generated exon database as previously described [15] with slight modification. Briefly, exon definitions were initially obtained from Ensembl release 88 (downloaded March 2017 containing 744,360 exons across 63,898 genes). This initial set was first condensed to remove redundancies (haplotypes, patches and alternate exon entries resulting from different source annotations) and exons from genes that were fully overlapping. Regions of exons containing overlaps belonging to different genes were removed; exons with at least 10 bp of sequence not contained within the overlapping region were retained by removing only the region of overlap, resulting in splitting up of some exons into one or more smaller unique exons. Exons for the same genes having overlaps of at least 1 bp were collapsed into single metaexons containing all overlapping exons. Human fasta sequences (for the reduced set of 320,753 meta exons across 56,479 genes) were then extracted from the genome (Ensembl GRCh38.p10). We then used BLAT (BLAT V. 35 [BLAT Citation]) to identify the orthologous exon sequences in the chimpanzee genome (panTro5). All hits with indels larger than 25 bp were removed (using a function blatOutIndelIdent, from [cite IGR bitbucket/github]) the fasta sequences for the

hits with the highest sequence identity were extracted from the genome (using Bedtools [citation]). The chimpanzee sequences were then blatted against the human genome and back to the chimpanzee genome. Queries that did not return the original blat locations (for either chimpanzee or human) were removed, as were entries where multiple possible mappings occurred with higher than 90% sequence identity. Finally, different exons mapping to the same location (different locations in human, but same location in chimpanzee) were removed. This updated human-chimpanzee orthologous exon database resulted in 254,172 metaexons across 44,125 genes.

4.4.6 Mapping of RNA-seq data, orthologous exons, read count transformation and normalization

We mapped human reads to the hg38 genome and chimpanzee reads to panTro5 using HISAT2 using the default parameters. We kept on only reads that mapped uniquely. To prevent biases in expression level estimates due to differences in mRNA transcript size and the relatively poor annotation of the chimpanzee genome and differences in mappability between species, we only kept reads that mapped to a list of orthologous metaexons. As the previous version of this metaexon database was generated using older versions of the genomes, we generated an updated orthoexon database based on the most current genome builds (see methods below for details). Gene expression levels were quantified using the feature counts function in SubRead. We performed all downstream processing and analysis steps in R unless otherwise stated. We transformed raw gene counts into $\log_2$-transformed counts per million (CPM) for each sample using edgeR [115]. To filter for the lowly expressed genes, we kept only genes with an average expression level of $\log_2$ (CPM) > 2 in at least one cell type [116]. For the remaining genes, we normalized the original read counts using the weighted trimmed mean of M-values algorithm (TMM) [116] to account for differences in the read counts at the extremes of the distribution and calculated the TMM-normalized $\log_2$-transformed CPM.

After TMM-normalization we then performed a cyclic loess normalization with the function normalizeCyclicLoess from the R/Bioconductor package limma [4, 113]. To account for gene length differences between species we calculated normalized $\log_2$-transformed RPKM values by using the function rpkm with normalized library sizes from the package edgeR [115]. We measured the gene lengths as the sum of the lengths of the orthologous exons. To partition the total gene expression variance into the variance attributed to each of the different properties of our study, we used a linear mixed model implemented in variancePartition [48]. We modeled all effects as fixed effects with the exception of individual which was modeled as a random effect.

4.4.7 Linear modeling, differential expression and GO analysis

Differential expression was estimated using a linear model based empirical Bayes method implemented in the R package limma [128, 129]. In order to use a linear modeling approach with RNA-seq read counts, we calculated weights that account for the mean-variance relationship of the count data using the function voom from the limma package [62]. In analyses with both tissue and iPSC derived cardiomyocytes, we cannot correct for purity differences, thus, we chose iPSC derived cardiomyocytes with the highest purity. Since RNA quality varied greatly across different tissues and derived cells (4.1), we modeled RNA quality as a fixed effect using RIN scores. For all pairwise differential expression comparisons, the species, cell type, RNA quality (RIN), and a species-by-cell type interaction were modeled as fixed effects, and individual as a random effect. We used contrast tests in limma to find genes that were differentially expressed between tissues and cells for each species. For each pairwise DE test, we corrected for multiple testing with the Benjamini & Hochberg false discovery rate [10] and genes with an FDR-adjusted P values < 0.05 were considered DE. To identify genes as shared or specific to one species when examining DE genes between different cell or tissue types while accounting for incomplete power to detect overlaps, we first identified DE

genes between the different cell or tissue types at an FDR of 5% in each species separately. Using this list of significantly DE genes, we then looked in the other species, but classified genes as DE with a more relaxed cut off using a nominal P value of 0.05. Genes identified at an FDR of 5% in one species but not significant at a nominal P value of less than 0.05 in the other species were determined to be specific to one species. Genes that were significant in one species at FDR of 5% and significant at a nominal P value of less than 0.05 in the other species were designated as shared. This same approach was also taken when looking at interspecies DE genes in heart tissue or iPSC derived cardiomyocytes. When examining the overlap of interspecies DE genes in multiple different tissues, data for each tissue or cell type was modeled separately, as all data could not be modeled together due differences in sequencing technology (50 bp Illumina GAIIx vs 100 bp Illumina HiSeq 4000). For this analysis, we only included terms for species and RNA quality (RIN score) (correcting for individual was not necessary in this case). To identify shared interspecies DE genes across different tissues and cell types while accounting for incomplete power to detect overlaps, we used the same approach as detailed above. To identify significantly enriched GO biological process terms for genes DE between day 27 cardiomyocytes and heart tissue, we used the compareCluster function implemented in the R package clusterProfiler (version 3.4.4, [170]). We used all expressed genes as our background for enrichment and only considered GO terms containing more than three annotated genes and less than 3000 genes. When considering enriched GO biological process terms for interspecies DE genes we used a binomial enrichment test with a Bonferroni correction for multiple testing as implemented in PantherDBs overrepresentation test [84, 85].

4.4.8 Determining how purity differences affect estimates of interspecies DE overlaps

Since target cell type heterogeneity may vary within tissues and cannot be easily assessed, we were interested in the effect of purity on our ability to detect interspecies tissue specific DE patterns using iPSC derived cardiomyocytes. As we are unable to physically alter (or effectively estimate) the composition of post mortem tissues, we instead used different combinations of day 27 cardiomyocyte samples generating different average purities of day 27 cardiomyocytes. With each combination of samples, we repeated the same analysis as described in the main text to determine how many interspecies DE genes identified in heart tissue could be recapitulated using iPSC derived cardiomyocytes.

4.5 Supplementary Information

4.5.1 Supplementary Figures

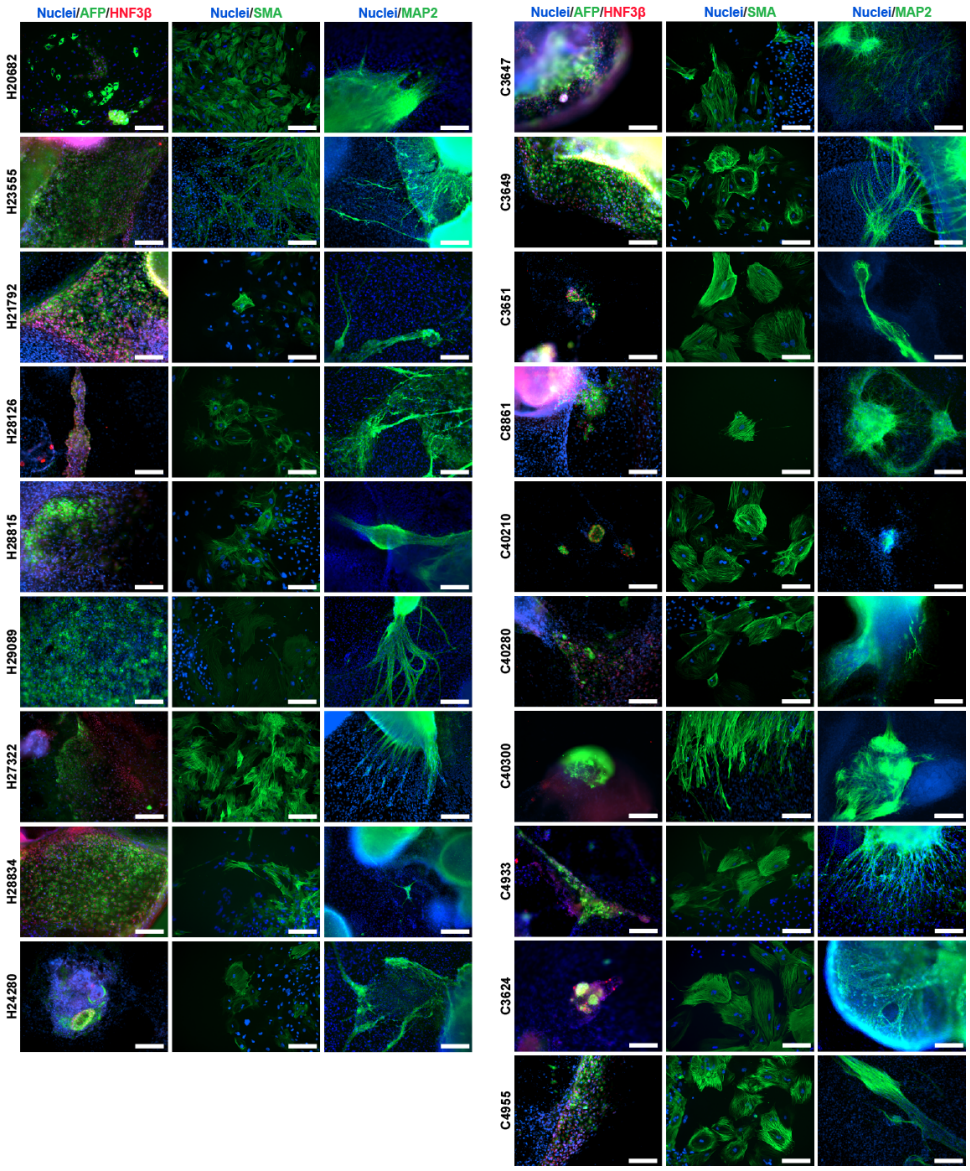


Figure 4.9: **Spontaneous differentiation assay of pluripotency for iPSC lines used in this study.** A. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for Human iPSC lines. B. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for chimpanzee iPSC lines, antibodies identifying celltypes derived from the three germ layers as indicated.Scale bar: 200 μm.

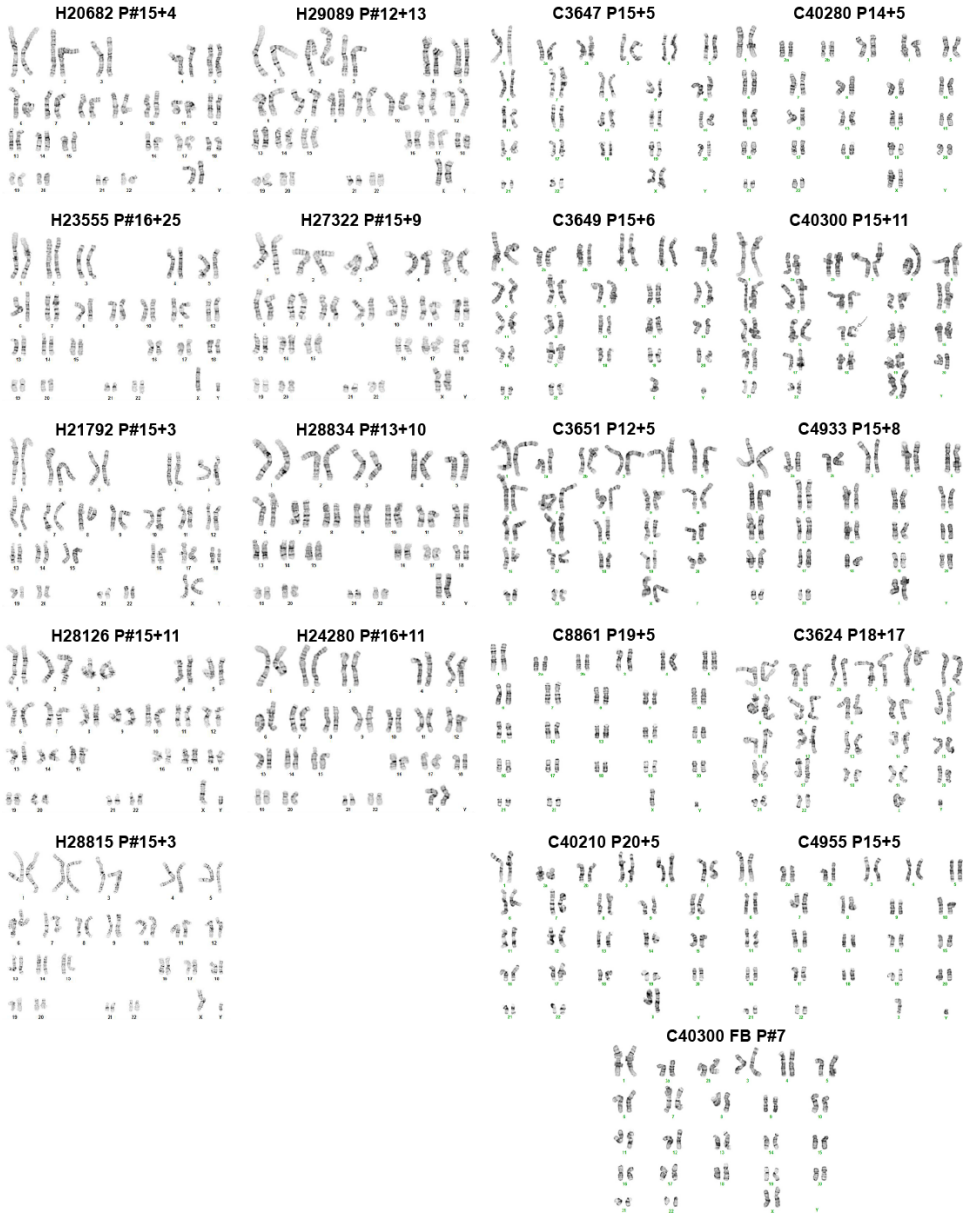


Figure 4.10: **Karyotypes for human and chimpanzee iPSC lines.** Karyotypes for human (right) and chimpanzee lines (left) used in this study. We identified additional bands in the p-arms of one chromosome 13 homolog and one chromosome 18 homolog for chimpanzee iPSC line C3. Thus we tested the source fibroblast line (C3 FB) to determine that these polymorphisms were normal polymorphisms and not formed de novo as a result of reprogramming.

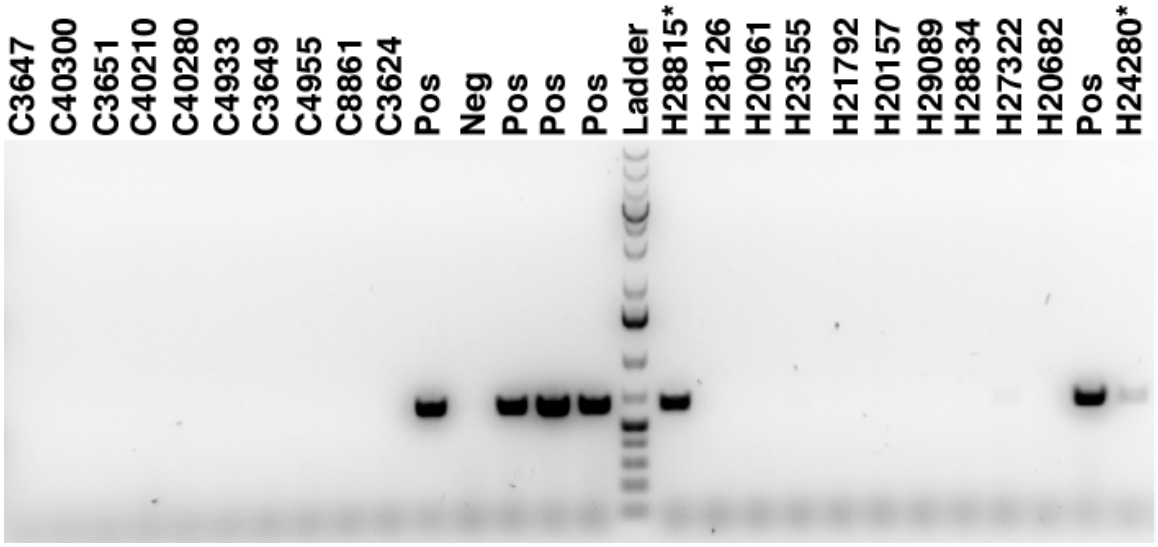


Figure 4.11: **PCR gel to test for exogenous episomal reprogramming vectors.** PCR gel for all iPSC lines used for this study. Pos indicates positive controls,Neg is a negative control. Human lines H28815 and H24280 demonstrate positive results for presence of reprogramming plasmid.

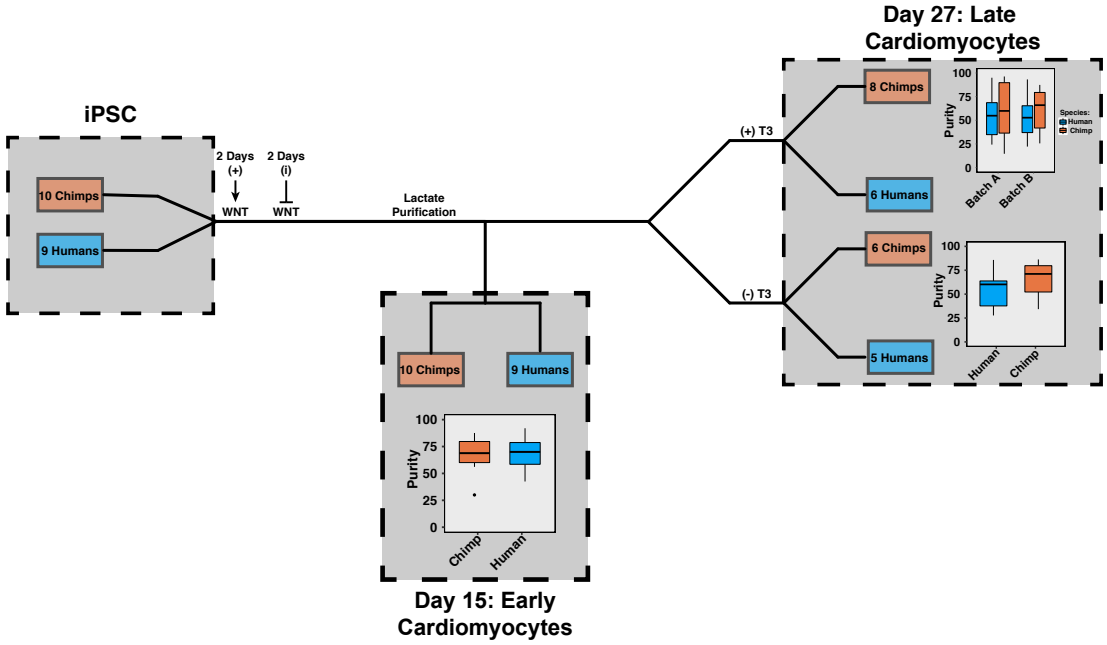


Figure 4.12: **Study design with iPSC-CM differentiation outline and summary of data collected.** Starting with iPSC lines we induced differentiation to iPSC-CMs using a previously published protocol using WNT agonist for 2 days followed by WNT antagonism for 2 days. At day 15 we harvested RNA from early iPSC-CMs after metabolic purification with lactic acid, within a single batch we collected 10 chimps and 9 human lines. Cells from a subset of individuals were replated and allowed to recover before starting treatment with T3 ((+)T3). A matched set of cells were cultured in parallel but not treated with T3 ((-)T3). We harvested RNA from (+)T3 and (-)T3 day 27 iPSC-CMs within the same batch that was matched to the day 15 collection (batch B). We also collected another set of (+)T3 treated Day 27 iPSC-CMs that did not have a matched (-)T3 or Day 15 set (denoted batch A). Purities obtained from dual positive (TNNI3+/TNNT2+) flow cytometry analysis for iPSC-CM samples are show in box plots. Grey shaded boxes denote where sample collection occurred.

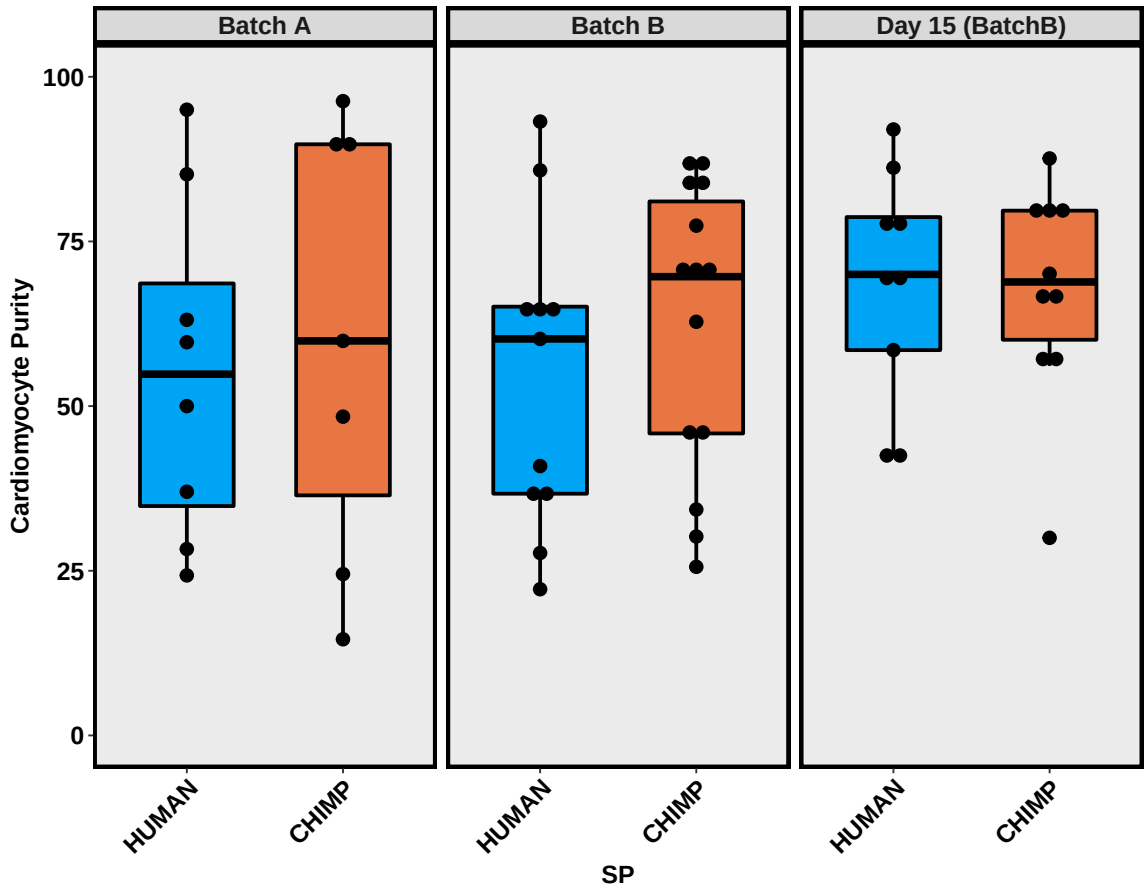


Figure 4.13: **Differentiation batch and purity are not associated with species.** Distribution of purities for human and chimpanzee iPSC derived cardiomyocytes for all samples collected in this study stratified by differentiation batch.

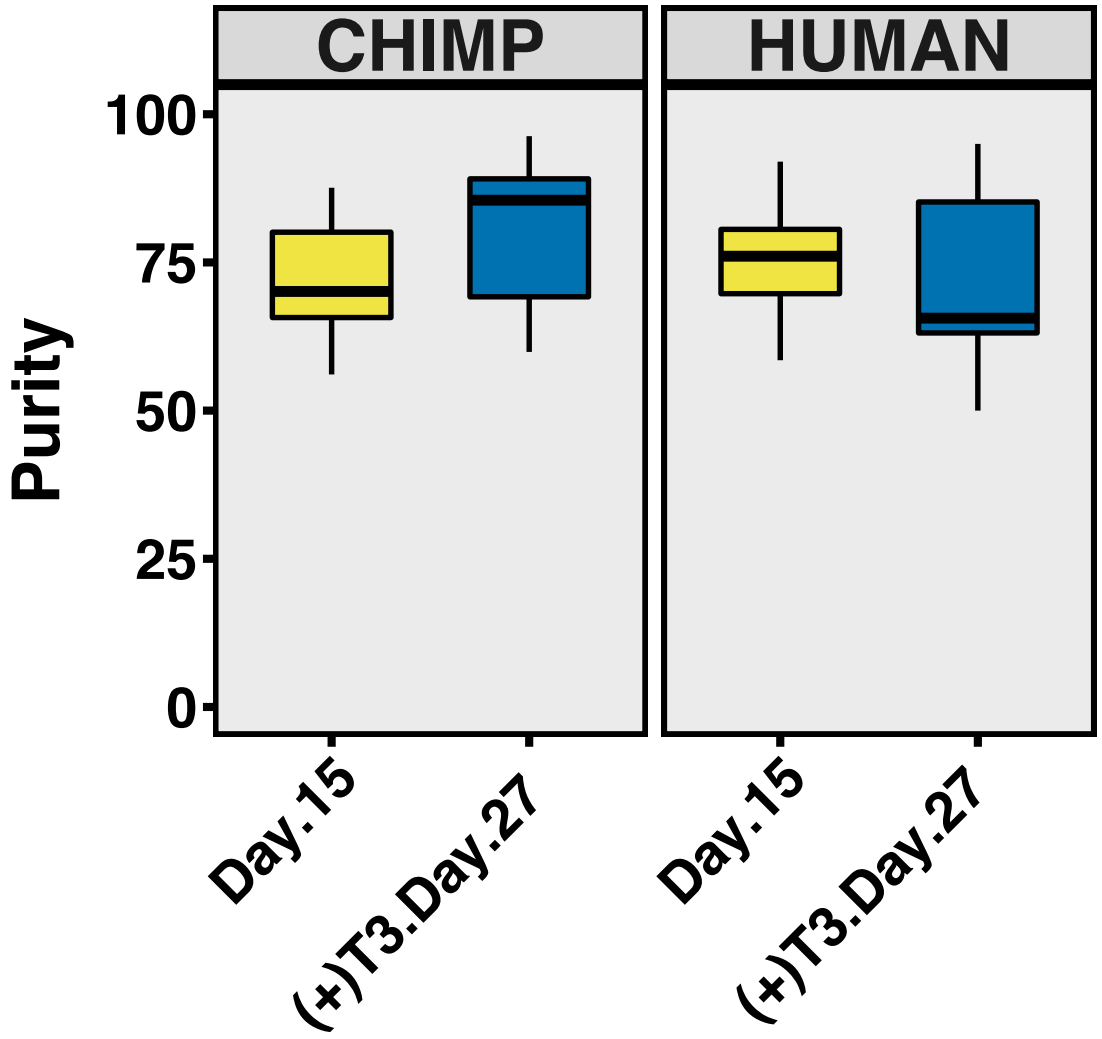


Figure 4.14: **Purity of samples used for analysis of genes that are under differential regulation within iPSCs, iPSC derived cardiomyocytes and heart tissue.**Box plots showing the purity of seven chimpanzee (left) and seven human (right) day 15 and day 27 iPSC derived cardiomyocytes used for analysis, see 4.1 for list of samples used.

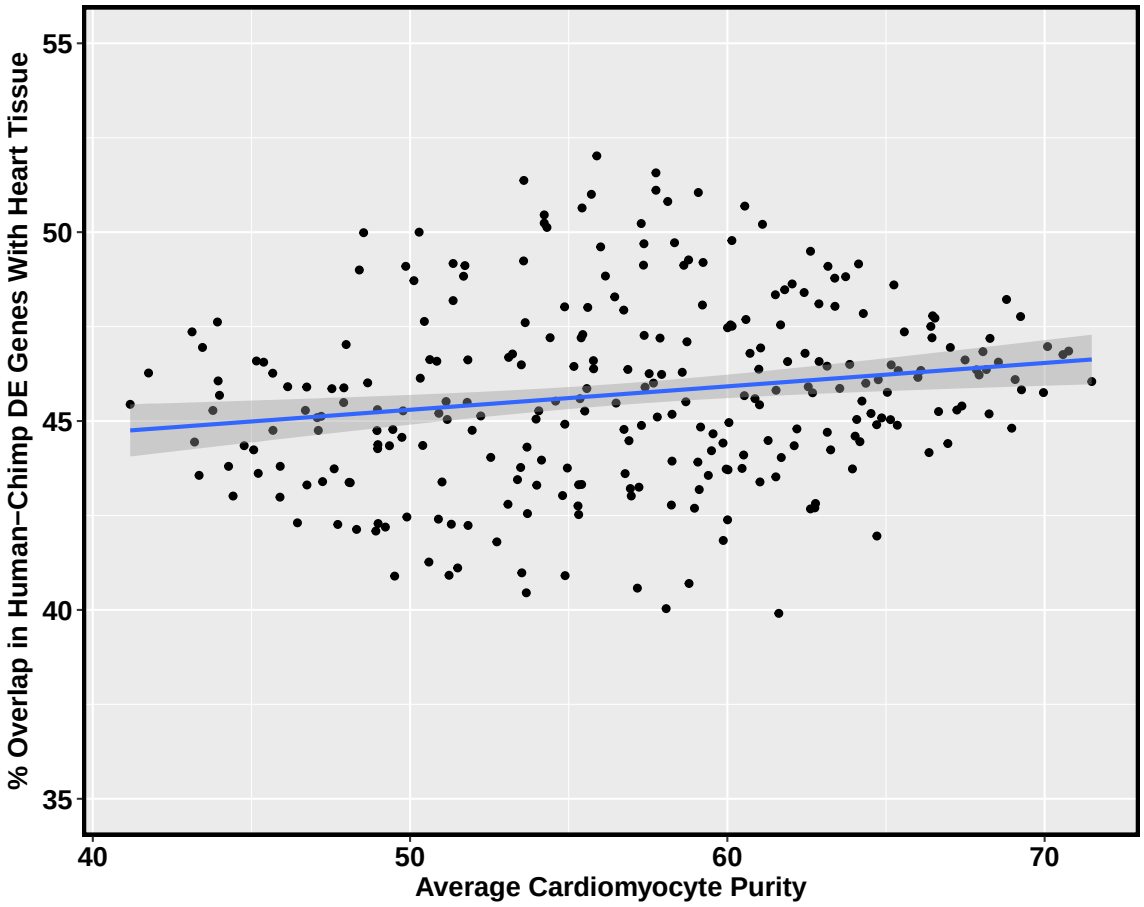


Figure 4.15: **Effect of purity on recapitulation if interspecies DE patterns using iPSC derived cardiomyocytes.** Scatter plot showing the average purity of sample set used on X axis and the percentage of human-chimp DE genes identified on the Y axis. Regression line is shown for Y X in blue with a 95% confidence interval in gray.

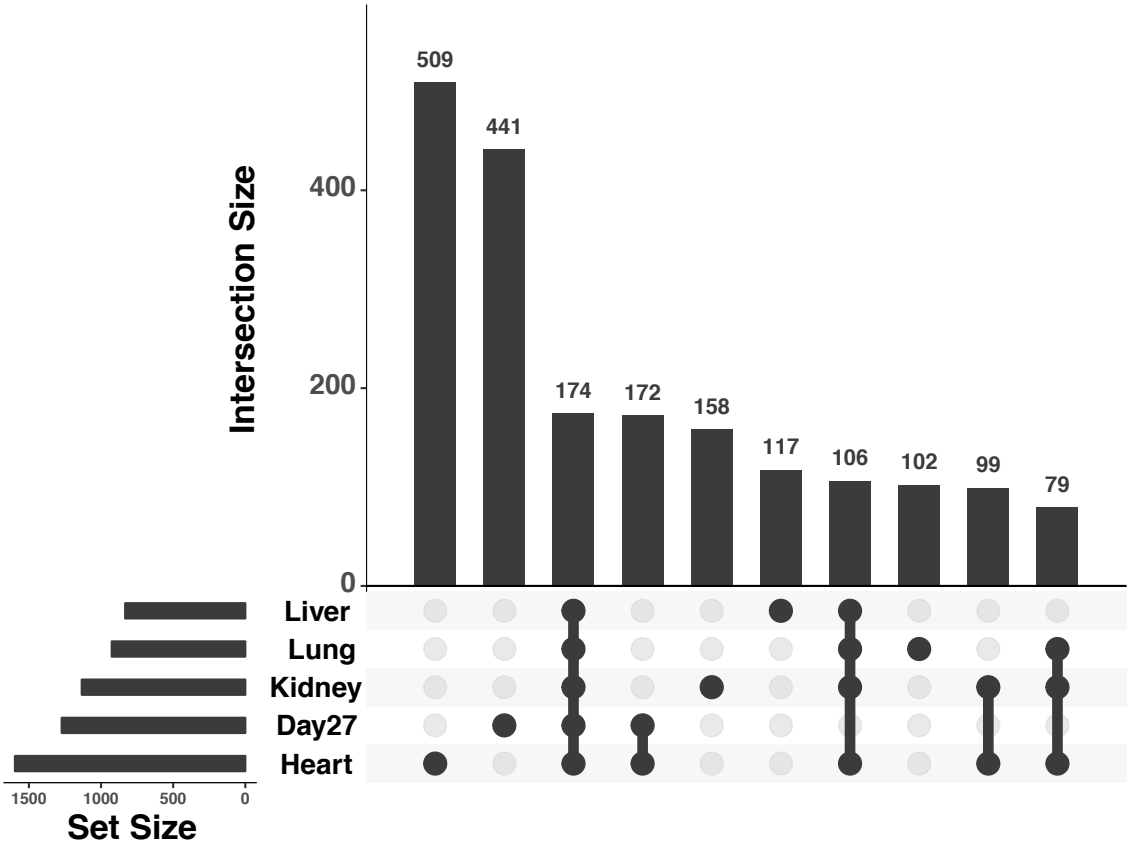


Figure 4.16: **Upsetr plot showing top 10 largest interspecies DE overlaps across multiple tissues and day 27 cardiomyocytes.**

4.5.2 Supplementary Tables

Table 4.1: **Sample meta data.**

Table 4.2: **Output summary for linear models of differentially expressed genes.**

Table 4.3: **GO results for genes differentially regulated between day 27 cardiomyocytes and heart tissues.**

Table 4.4: **GO results for genes differentially regulated between human and chimpanzeee in day 27 cardiomyocytes and heart tissues.**

CHAPTER 5

CONCLUSION

In this dissertation, I present work that has focused on the goal of developing a novel cellular novel system to conduct comparative primate genomics studies. To date most comparative primate genomics studies have been conducted using a small collection of primate cell lines (almost exclusively fibroblasts and LCLs) or using post mortem tissues. These studies have yielded many insights into the regulatory differences distinguishing humans from our closest living relative, chimpanzees (pan troglodytes). However, post-mortem tissues are a static nonrenewable resource associated with high environmental variances, while the collection of cell lines available from chimpanzee and other primates is fairly limited, thus restricting our ability to examine the various gene regulatory contexts provided by differing cellular populations. In order to circumvent these limitations in chapter 2, I establish a novel resource to conduct comparative primate genomic studies. Harnessing the power of IPSC technology I generated a matched panel of pluripotent stem cells from both human and chimpanzee fibroblasts using the same protocol of generation. To demonstrate the utility of this panel for comparative genomics, we collected RNA-sequencing, methylation and CHIP-seq for both the iPSC lines and the somatic cell lines used to generate each line. Globally we found that within iPSCs, individuals within each species are more similar to each other than they are in the mature somatic cell types that the iPSCs were derived from. When we compare across species within iPSCs, I identified 4500 genes that were differentially expressed, and over 3500 genomic regions with different patterns of DNA methylation. While this seems like a surprising number of differences given the expectations set by our close evolutionary relationship and previous studies conducted in tissues. The majority of DE genes have moderate fold-change differences which we likely had increased power to detect given the reduced inter-individual variance observed within iPSCs. While this newly created resource represents a valuable resource and a large step forward for genomic studies, iPSC lines

only provide a snapshot of the possible cellular regulatory states found within human and chimpanzees. To access the remaining cellular states requires that we induce different ion using protocols that have been developed and guided by the principles of development. Often times these protocols were optimized and tailored to generate high cellular yields of target cell types for humans (or in some cases mice). Therefore, it is reasonable that, as a result, the ensuing in vitro process may entail a departure from the canalized developmental process occurring in vivo for each species. Therefore, in chapter 3, I set out to assay how human and chimpanzee iPSCs are able to recapitulate a dynamic developmental process with expected conservation across species. This was achieved by examining gene expression dynamics (using RNA-sew) during a four-day time course differentiation of human and chimpanzee iPSCs into endoderm. By jointly analyzing the data from the entire time course, I found that nearly all patterns (75%) of gene expression trajectories, are shared across the two species. Furthermore, I observed a marked reduction of both intra- and inter-species variation in gene expression levels at the initial stages of differentiation (primitive streak) compared to the iPSCs. While the results in chapter 3 suggest that both species are progressing through a similar canalized developmental process. There remains a possibility that differentiated cells derived from primate iPSCs do not perfectly recapitulate gene regulatory differences observed within primary tissues between species. Thus, in chapter 4, I compared the global gene regulatory state of iPSC derived cardiomyocytes from both human and chimpanzees with that of primary heart tissues from each species. I was able to identify the optimal parameters of iPSC differentiation into cardiomyocytes that minimized differences between each species and their respective adult tissue counterparts. Furthermore, I determined that globally, cultured cardiomyocytes appear more similar to adult hearts than any other tissue. Finally, I found that iPSC derived cardiomyocytes are able to recapitulate up to 50% of the interspecies gene regulatory differences identified in primary post mortem heart tissues. The work contained in this thesis represents an important step forward in the development of

new tools for comparative studies of primate gene regulation. In order to further develop these tools and their utility in comparative studies, our lab provides the iPSC lines developed herein freely and openly with fellow researchers, to date the panel has already been shared with over 20 laboratories.

References

[1] et al. Aguet F, Brown AA. Local genetic effects on gene expression across 44 human tissues. *bioRxiv*, 2016.

[2] A. Alexa, J. Rahnenfuhrer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–7, 2006.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000.

[4] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau. Faster cyclic loess: normalizing rna arrays via linear models. *Bioinformatics*, 20(16):2778–86, 2004.

[5] Nicholas E. Banovich, Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E. Burnett, Marsha Myrthil, Samantha M. Thomas, Courtney K. Burrows, Irene Gallego Romero, Bryan J. Pavlovic, Anshul Kundaje, Jonathan K. Pritchard, and Yoav Gilad. Impact of regulatory variation across human ipscs and differentiated cells. *bioRxiv*, page doi: 10.1101.091660, 2016.

[6] L. B. Barreiro, J. C. Marioni, R. Blekhman, M. Stephens, and Y. Gilad. Functional comparison of innate immune signaling pathways in primates. *PLoS Genet*, 6(12):e1001249, 2010.

[7] A. M. Barrette, J. K. Roberts, C. Chapin, E. A. Egan, M. R. Segal, J. A. Oses-Prieto, S. Chand, A. L. Burlingame, and P. L. Ballard. Antiinflammatory effects of budesonide in human fetal lung. *Am J Respir Cell Mol Biol*, 55(5):623–632, 2016.

[8] C. C. Bell, V. M. Lauschke, S. U. Vorrink, H. Palmgren, R. Duffin, T. B. Andersson, and M. Ingelman-Sundberg. Transcriptional, functional, and mechanistic comparisons of stem cell-derived hepatocytes, heparg cells, and three-dimensional human hepatocyte spheroids as predictive in vitro systems for drug-induced liver injury. *Drug Metab Dispos*, 45(4):419–429, 2017.

[9] I. F. Ben-Nun, S. C. Montague, M. L. Houck, H. T. Tran, I. Garitaonandia, T. R. Leonardo, Y. C. Wang, S. J. Charter, L. C. Laurent, O. A. Ryder, and J. F. Loring. Induced pluripotent stem cells from highly endangered species. *Nat Methods*, 8(10):829–31, 2011.

[10] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[11] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–26, 2006.

[12] Lauren E. Blake, Samantha M. Thomas, John D. Blischak, Chiaowen Joyce Hsiao, Claudia Chavarria, Marsha Myrthil, Yoav Gilad, and Bryan J. Pavlovic. A comparative study of endoderm differentiation in humans and chimpanzees. *bioRxiv*, 2017.

[13] R. Blekhman, J. C. Marioni, P. Zumbo, M. Stephens, and Y. Gilad. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res*, 20(2):180–9, 2010.

[14] R. Blekhman, A. Oshlack, A. E. Chabot, G. K. Smyth, and Y. Gilad. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet*, 4(11):e1000271, 2008.

[15] Ran Blekhman. A database of orthologous exons in primates for comparative analysis of rna-seq data. *Nature Precedings*, 2012.

[16] J. D. Blischak, L. Tailleux, A. Mitrano, L. B. Barreiro, and Y. Gilad. Mycobacterial infection induces a specific human innate immune response. *Sci Rep*, 5:16882, 2015.

[17] C. Bock, E. Kiskinis, G. Verstappen, H. Gu, G. Boulting, Z. D. Smith, M. Ziller, G. F. Croft, M. W. Amoroso, D. H. Oakley, A. Gnirke, K. Eggan, and A. Meissner. Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144(3):439–52, 2011.

[18] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grutzner, S. Bergmann, R. Nielsen, S. Paabo, and H. Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–8, 2011.

[19] Roy J Britten and Eric H Davidson. Repetitive and non-repetitive dna sequences and a specualation on the origins of evolutionary novelty. *The Quarterly Review of Biology*, 46:111–138, 1971.

[20] A. H. Brivanlou, F. H. Gage, R. Jaenisch, T. Jessell, D. Melton, and J. Rossant. Stem cells. setting standards for human embryonic stem cells. *Science*, 300(5621):913–6, 2003.

[21] P. W. Burridge, A. Holmstrom, and J. C. Wu. Chemically defined culture and cardiomyocyte differentiation of human pluripotent stem cells. *Curr Protoc Hum Genet*, 87:21 3 1–15, 2015.

[22] P. W. Burridge, E. Matsa, P. Shukla, Z. C. Lin, J. M. Churko, A. D. Ebert, F. Lan, S. Diecke, B. Huber, N. M. Mordwinkin, J. R. Plews, O. J. Abilez, B. Cui, J. D. Gold, and J. C. Wu. Chemically defined generation of human cardiomyocytes. *Nat Methods*, 11(8):855–60, 2014.

[23] C. E. Cain, R. Blekhman, J. C. Marioni, and Y. Gilad. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*, 187(4):1225–34, 2011.

[24] S. B. Carroll. Evolution at two levels: on genes and form. *PLoS Biol*, 3(7):e245, 2005.

[25] S. B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008.

[26] A. W. Chan, P. H. Cheng, A. Neumann, and J. J. Yang. Reprogramming huntington monkey skin cells into pluripotent stem cells. *Cell Reprogram*, 12(5):509–17, 2010.

[27] G. Chen, D. R. Gulbranson, Z. Hou, J. M. Bolin, V. Ruotti, M. D. Probasco, K. Smuga-Otto, S. E. Howden, N. R. Diol, N. E. Propson, R. Wagner, G. O. Lee, J. Antosiewicz-Bourget, J. M. Teng, and J. A. Thomson. Chemically defined conditions for human ipsc derivation and culture. *Nat Methods*, 8(5):424–9, 2011.

[28] K. G. Chen, B. S. Mallon, R. S. Hamilton, O. A. Kozhich, K. Park, D. J. Hoeppner, P. G. Robey, and R. D. McKay. Non-colony type monolayer culture of human embryonic stem cells. *Stem Cell Res*, 9(3):237–48, 2012.

[29] X. Cheng, L. Ying, L. Lu, A. M. Galvao, J. A. Mills, H. C. Lin, D. N. Kotton, S. S. Shen, M. C. Nostro, J. K. Choi, M. J. Weiss, D. L. French, and P. Gadue. Self-renewing endodermal progenitor lines generated from human pluripotent stem cells. *Cell Stem Cell*, 10(4):371–84, 2012.

[30] Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Gallego Romero I, Pritchard JK, and Gilad Y. Genetic variation, not cell type of origin, underlies the majority of identifiable regulatory differences in ipscs. *PloS Genetics*, 12(1):e1005793. doi: 10.1371/journal.pgen.1005793, 2016.

[31] H. Clevers. Modeling development and disease with organoids. *Cell*, 165(7):1586–1597, 2016.

[32] D. E. Cohen and D. Melton. Turning straw into gold: directing cell fate for regenerative medicine. *Nat Rev Genet*, 12(4):243–52, 2011.

[33] J. R. Conway, A. Lex, and N. Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

[34] J. Cotney, J. Leng, J. Yin, S. K. Reilly, L. E. DeMare, D. Emera, A. E. Ayoub, P. Rakic, and J. P. Noonan. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, 154(1):185–96, 2013.

[35] K. A. D'Amour, A. D. Agulnick, S. Eliazer, O. G. Kelly, E. Kroon, and E. E. Baetge. Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat Biotechnol*, 23(12):1534–41, 2005.

[36] M. Deleidi, G. Hargus, P. Hallett, T. Osborn, and O. Isacson. Development of histocompatible primate-induced pluripotent stem cells for neural transplantation. *Stem Cells*, 29(7):1052–63, 2011.

[37] P. Du, W. A. Kibbe, and S. M. Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, 24(13):1547–8, 2008.

[38] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, 2004.

[39] R. Ellen Kreipke, Y. Wang, J. W. Miklas, J. Mathieu, and H. Ruohola-Baker. Metabolic remodeling in early development and cardiomyocyte maturation. *Semin Cell Dev Biol*, 52:84–92, 2016.

[40] W. Enard, S. Gehre, K. Hammerschmidt, S. M. Holter, T. Blass, M. Somel, M. K. Bruckner, C. Schreiweis, C. Winter, R. Sohr, L. Becker, V. Wiebe, B. Nickel, T. Giger, U. Muller, M. Groszer, T. Adler, A. Aguilar, I. Bolle, J. Calzada-Wack, C. Dalke, N. Ehrhardt, J. Favor, H. Fuchs, V. Gailus-Durner, W. Hans, G. Holzlwimmer, A. Javaheri, S. Kalaydjiev, M. Kallnik, E. Kling, S. Kunder, I. Mossbrugger, B. Naton, I. Racz, B. Rathkolb, J. Rozman, A. Schrewe, D. H. Busch, J. Graw, B. Ivandic, M. Klingenspor, T. Klopstock, M. Ollert, L. Quintanilla-Martinez, H. Schulz, E. Wolf, W. Wurst, A. Zimmer, S. E. Fisher, R. Morgenstern, T. Arendt, M. H. de Angelis, J. Fischer, J. Schwarz, and S. Paabo. A humanized version of foxp2 affects cortico-basal ganglia circuits in mice. *Cell*, 137(5):961–71, 2009.

[41] T. Faial, A. S. Bernardo, S. Mendjan, E. Diamanti, D. Ortmann, G. E. Gentsch, V. L. Mascetti, M. W. Trotter, J. C. Smith, and R. A. Pedersen. Brachyury and smad signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development*, 142(12):2121–35, 2015.

[42] I. Gallego Romero, B. J. Pavlovic, I. Hernando-Herraez, X. Zhou, M. C. Ward, N. E. Banovich, C. L. Kagan, J. E. Burnett, C. H. Huang, A. Mitrano, C. I. Chavarria, I. Friedrich Ben-Nun, Y. Li, K. Sabatini, T. R. Leonardo, M. Parast, T. Marques-Bonet, L. C. Laurent, J. F. Loring, and Y. Gilad. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *Elife*, 4:e07103, 2015.

[43] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

[44] Y. Gilad, S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res*, 15(5):674–80, 2005.

[45] A. E. Handel, S. Chintawar, T. Lalic, E. Whiteley, J. Vowles, A. Giustacchini, K. Argoud, P. Sopp, M. Nakanishi, R. Bowden, S. Cowley, S. Newey, C. Akerman, C. P. Ponting, and M. Z. Cader. Assessing similarity to primary tissue and cortical layer identity in induced pluripotent stem cell-derived cortical neurons through single-cell transcriptomics. *Hum Mol Genet*, 25(5):989–1000, 2016.

[46] C. He, H. Hu, K. D. Wilson, H. Wu, J. Feng, S. Xia, J. Churko, K. Qu, H. Y. Chang, and J. C. Wu. Systematic characterization of long noncoding rnas reveals the contrasting coordination of cis- and trans-molecular regulation in human fetal and adult hearts. *Circ Cardiovasc Genet*, 9(2):110–8, 2016.

[47] I. Hernando-Herraez, J. Prado-Martinez, P. Garg, M. Fernandez-Callejo, H. Heyn, C. Hvilsom, A. Navarro, M. Esteller, A. J. Sharp, and T. Marques-Bonet. Dynamics of dna methylation in recent human and great ape evolution. *PLoS Genet*, 9(9):e1003763, 2013.

[48] G. E. Hoffman and E. E. Schadt. variancepartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, 17(1):483, 2016.

[49] S. G. Hong, T. Winkler, C. Wu, V. Guo, S. Pittaluga, A. Nicolae, R. E. Donahue, M. E. Metzger, S. D. Price, N. Uchida, S. A. Kuznetsov, T. Kilts, L. Li, P. G. Robey, and C. E. Dunbar. Path to the clinic: assessment of ipsc-based cell therapies in vivo in a nonhuman primate model. *Cell Rep*, 7(4):1298–1309, 2014.

[50] S. E. Howden, H. Wardan, L. Voullaire, S. McLenachan, R. Williamson, P. Ioannou, and J. Vadolas. Chromatin-binding regions of ebna1 protein facilitate the enhanced transfection of epstein-barr virus-based vectors. *Hum Gene Ther*, 17(8):833–44, 2006.

[51] M. A. Israel, S. H. Yuan, C. Bardy, S. M. Reyna, Y. Mu, C. Herrera, M. P. Hefferan, S. Van Gorp, K. L. Nazor, F. S. Boscolo, C. T. Carson, L. C. Laurent, M. Marsala, F. H. Gage, A. M. Remes, E. H. Koo, and L. S. Goldstein. Probing sporadic and familial alzheimer's disease using induced pluripotent stem cells. *Nature*, 482(7384):216–20, 2012.

[52] C. Y. Ivashchenko, G. C. Pipes, I. M. Lozinskaya, Z. Lin, X. Xiaoping, S. Needle, E. T. Grygielko, E. Hu, J. R. Toomey, J. J. Lepore, and R. N. Willette. Human-induced pluripotent stem cell-derived cardiomyocytes exhibit temporal changes in phenotype. *Am J Physiol Heart Circ Physiol*, 305(6):H913–22, 2013.

[53] F. Jacob. Evolution and tinkering. *Science*, 196(4295):1161–6, 1977.

[54] R. Jaenisch and R. Young. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, 132(4):567–82, 2008.

[55] S. J. Kattman, A. D. Witty, M. Gagliardi, N. C. Dubois, M. Niapour, A. Hotta, J. Ellis, and G. Keller. Stage-specific optimization of activin/nodal and bmp signaling promotes cardiac differentiation of mouse and human pluripotent stem cell lines. *Cell Stem Cell*, 8(2):228–40, 2011.

[56] S. W. Kembel, P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–4, 2010.

[57] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36, 2013.

[58] Mary-Claire King and A. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188:107–116, 1975.

[59] T. D. Kocher, W. K. Thomas, A. Meyer, S. V. Edwards, S. Paabo, F. X. Villablanca, and A. C. Wilson. Dynamics of mitochondrial dna evolution in animals: amplification and sequencing with conserved primers. *Proc Natl Acad Sci U S A*, 86(16):6196–200, 1989.

[60] T. J. Kolanowski, C. L. Antos, and K. Guan. Making human cardiomyocytes up to date: Derivation, maturation state and perspectives. *Int J Cardiol*, 241:379–386, 2017.

[61] A. Kyttala, R. Moraghebi, C. Valensisi, J. Kettunen, C. Andrus, K. K. Pasumarthy, M. Nakanishi, K. Nishimura, M. Ohtaka, J. Weltner, B. Van Handel, O. Parkkonen, J. Sinisalo, A. Jalanko, R. D. Hawkins, N. B. Woods, T. Otonkoski, and R. Trokovic. Genetic variability overrides the impact of parental cell type and determines ipsc differentiation potential. *Stem Cell Reports*, 6(2):200–12, 2016.

[62] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

[63] Y. K. Lee, K. M. Ng, Y. C. Chan, W. H. Lai, K. W. Au, C. Y. Ho, L. Y. Wong, C. P. Lau, H. F. Tse, and C. W. Siu. Triiodothyronine promotes cardiac differentiation and maturation of embryonic stem cells via the classical genomic pathway. *Mol Endocrinol*, 24(9):1728–36, 2010.

[64] G. Li, A. Xu, S. Sim, J. R. Priest, X. Tian, T. Khan, T. Quertermous, B. Zhou, P. S. Tsao, S. R. Quake, and S. M. Wu. Transcriptomic profiling maps anatomically patterned subpopulations among single embryonic cardiac cells. *Dev Cell*, 39(4):491–507, 2016.

[65] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[66] Q. Li, S. Lian, Z. Dai, Q. Xiang, and X. Dai. Bgdb: a database of bivalent genes. *Database (Oxford)*, 2013:bat057, 2013.

[67] X. Lian, J. Zhang, S. M. Azarin, K. Zhu, L. B. Hazeltine, X. Bao, C. Hsiao, T. J. Kamp, and S. P. Palecek. Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating wnt/beta-catenin signaling under fully defined conditions. *Nat Protoc*, 8(1):162–75, 2013.

[68] Y. Liao, G. K. Smyth, and W. Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–30, 2014.

[69] B. Lin, X. Lin, M. Stachel, E. Wang, Y. Luo, J. Lader, X. Sun, M. Delmar, and L. Bu. Culture in glucose-depleted medium supplemented with fatty acid and 3,3',5-triiodo-l-thyronine facilitates purification and maturation of human pluripotent stem cell-derived cardiomyocytes. *Front Endocrinol (Lausanne)*, 8:253, 2017.

[70] S. Lin, Y. Lin, J. R. Nery, M. A. Urich, A. Breschi, C. A. Davis, A. Dobin, C. Zaleski, M. A. Beer, W. C. Chapman, T. R. Gingeras, J. R. Ecker, and M. P. Snyder. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A*, 111(48):17224–9, 2014.

[71] H. Liu, F. Zhu, J. Yong, P. Zhang, P. Hou, H. Li, W. Jiang, J. Cai, M. Liu, K. Cui, X. Qu, T. Xiang, D. Lu, X. Chi, G. Gao, W. Ji, M. Ding, and H. Deng. Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. *Cell Stem Cell*, 3(6):587–90, 2008.

[72] J. Liu, K. A. Koscielska, Z. Cao, S. Hulsizer, N. Grace, G. Mitchell, C. Nacey, J. Githinji, J. McGee, D. Garcia-Arocena, R. J. Hagerman, J. Nolta, I. N. Pessah, and P. J. Hagerman. Signaling defects in ipsc-derived fragile x premutation neurons. *Hum Mol Genet*, 21(17):3795–805, 2012.

[73] K. M. Loh, L. T. Ang, J. Zhang, V. Kumar, J. Ang, J. Q. Auyeong, K. L. Lee, S. H. Choo, C. Y. Lim, M. Nichane, J. Tan, M. S. Noghabi, L. Azzola, E. S. Ng, J. Durruthy-Durruthy, V. Sebastiano, L. Poellinger, A. G. Elefanty, E. G. Stanley, Q. Chen, S. Prabhakar, I. L. Weissman, and B. Lim. Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell*, 14(2):237–52, 2014.

[74] R. J. Lund, E. Narva, and R. Lahesmaa. Genetic and epigenetic stability of human pluripotent stem cells. *Nat Rev Genet*, 13(10):732–44, 2012.

[75] H. Ma, R. Morey, R. C. O'Neil, Y. He, B. Daughtry, M. D. Schultz, M. Hariharan, J. R. Nery, R. Castanon, K. Sabatini, R. D. Thiagarajan, M. Tachibana, E. Kang, R. Tippner-Hedges, R. Ahmed, N. M. Gutierrez, C. Van Dyken, A. Polat, A. Sugawara, M. Sparman, S. Gokhale, P. Amato, D. P. Wolf, J. R. Ecker, L. C. Laurent, and S. Mitalipov. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature*, 511(7508):177–83, 2014.

[76] M. C. N. Marchetto, I. Narvaiza, A. M. Denli, C. Benner, T. A. Lazzarini, J. L. Nathanson, A. C. M. Paquola, K. N. Desai, R. H. Herai, M. D. Weitzman, G. W. Yeo,

A. R. Muotri, and F. H. Gage. Differential l1 regulation in pluripotent stem cells of humans and apes. *Nature*, 503(7477):525–529, 2013.

[77] Maria C N Marchetto, Alysson R Muotri, and Fred H Gage. Proposing a model for studying primate development using induced pluripotent stem cells. *Programmed Cells from Basic Neuroscience to Therapy*, 20:31–39, 2013.

[78] P. A. Marinho, T. Chailangkarn, and A. R. Muotri. Systematic optimization of human pluripotent stem cells media using design of experiments. *Sci Rep*, 5:9834, 2015.

[79] S. Masui, Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. Ko, and H. Niwa. Pluripotency governed by sox2 via regulation of oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol*, 9(6):625–35, 2007.

[80] S. Masui, S. Ohtsuka, R. Yagi, K. Takahashi, M. S. Ko, and H. Niwa. Rex1/zfp42 is dispensable for pluripotency in mouse es cells. *BMC Dev Biol*, 8:45, 2008.

[81] C. Y. McLean, P. L. Reno, A. A. Pollen, A. I. Bassan, T. D. Capellini, C. Guenther, V. B. Indjeian, X. Lim, D. B. Menke, B. T. Schaar, A. M. Wenger, G. Bejerano, and D. M. Kingsley. Human-specific loss of regulatory dna and the evolution of human-specific traits. *Nature*, 471(7337):216–9, 2011.

[82] L. F. Meisner and J. A. Johnson. Protocols for cytogenetic studies of human embryonic stem cells. *Methods*, 45(2):133–41, 2008.

[83] F. T. Merkle and K. Eggan. Modeling human disease with pluripotent stem cells: from genome association to function. *Cell Stem Cell*, 12(6):656–68, 2013.

[84] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas. Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*, 45(D1):D183–D189, 2017.

[85] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas. Large-scale gene function analysis with the panther classification system. *Nat Protoc*, 8(8):1551–66, 2013.

[86] F. Mora-Bermudez, F. Badsha, S. Kanton, J. G. Camp, B. Vernot, K. Kohler, B. Voigt, K. Okita, T. Maricic, Z. He, R. Lachmann, S. Paabo, B. Treutlein, and W. B. Huttner. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife*, 5:e18386. doi: 10.7554/eLife.18683, 2016.

[87] F. J. Muller, B. M. Schuldt, R. Williams, D. Mason, G. Altun, E. P. Papapetrou, S. Danner, J. E. Goldmann, A. Herbst, N. O. Schmidt, J. B. Aldenhoff, L. C. Laurent, and J. F. Loring. A bioinformatic assay for pluripotency in human cells. *Nat Methods*, 8(4):315–7, 2011.

[88] C. E. Murry and G. Keller. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell*, 132(4):661–80, 2008.

[89] E. J. Nazareth, J. E. Ostblom, P. B. Lucker, S. Shukla, M. M. Alvarez, S. K. Oh, T. Yin, and P. W. Zandstra. High-throughput fingerprinting of human pluripotent stem cell fate responses and lineage bias. *Nat Methods*, 10(12):1225–31, 2013.

[90] K. L. Nazor, G. Altun, C. Lynch, H. Tran, J. V. Harness, I. Slavin, I. Garitaonandia, F. J. Muller, Y. C. Wang, F. S. Boscolo, E. Fakunle, B. Dumevska, S. Lee, H. S. Park, T. Olee, D. D. D'Lima, R. Semechkin, M. M. Parast, V. Galat, A. L. Laslett, U. Schmidt, H. S. Keirstead, J. F. Loring, and L. C. Laurent. Recurrent variations in dna methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell*, 10(5):620–34, 2012.

[91] H. H. Ng and M. A. Surani. The transcriptional and signalling networks of pluripotency. *Nat Cell Biol*, 13(5):490–6, 2011.

[92] K. Nowick, T. Gernat, E. Almaas, and L. Stubbs. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A*, 106(52):22358–63, 2009.

[93] S. S. Nunes, J. W. Miklas, J. Liu, R. Aschar-Sobbi, Y. Xiao, B. Zhang, J. Jiang, S. Masse, M. Gagliardi, A. Hsieh, N. Thavandiran, M. A. Laflamme, K. Nanthakumar, G. J. Gross, P. H. Backx, G. Keller, and M. Radisic. Biowire: a platform for maturation of human pluripotent stem cell-derived cardiomyocytes. *Nat Methods*, 10(8):781–7, 2013.

[94] X. Nuttle, G. Giannuzzi, M. H. Duyzend, J. G. Schraiber, I. Narvaiza, P. H. Sudmant, O. Penn, G. Chiatante, M. Malig, J. Huddleston, C. Benner, F. Camponeschi, S. Ciofi-Baffoni, H. A. Stessman, M. C. Marchetto, L. Denman, L. Harshman, C. Baker, A. Raja, K. Penewit, N. Janke, W. J. Tang, M. Ventura, L. Banci, F. Antonacci, J. M. Akey, C. T. Amemiya, F. H. Gage, A. Reymond, and E. E. Eichler. Emergence of a homo sapiens-specific gene family and chromosome 16p11.2 cnv susceptibility. *Nature*, 536(7615):205–9, 2016.

[95] S. Ogawa, J. Surapisitchat, C. Virtanen, M. Ogawa, M. Niapour, K. S. Sugamori, S. Wang, L. Tamblyn, C. Guillemette, E. Hoffmann, B. Zhao, S. Strom, R. R. Laposa, R. F. Tyndale, D. M. Grant, and G. Keller. Three-dimensional culture and camp signaling promote the maturation of human pluripotent stem cell-derived hepatocytes. *Development*, 140(15):3285–96, 2013.

[96] S. Okamoto and M. Takahashi. Induction of retinal pigment epithelial cells from monkey ips cells. *Invest Ophthalmol Vis Sci*, 52(12):8785–90, 2011.

[97] K. Okita, Y. Matsumura, Y. Sato, A. Okada, A. Morizane, S. Okamoto, H. Hong, M. Nakagawa, K. Tanabe, K. Tezuka, T. Shibata, T. Kunisada, M. Takahashi, J. Taka-

hashi, H. Saji, and S. Yamanaka. A more efficient method to generate integration-free human ips cells. *Nat Methods*, 8(5):409–12, 2011.

[98] S. H. Orkin and K. Hochedlinger. Chromatin connections to pluripotency and cellular reprogramming. *Cell*, 145(6):835–50, 2011.

[99] K. Osafune, L. Caron, M. Borowiak, R. J. Martinez, C. S. Fitz-Gerald, Y. Sato, C. A. Cowan, K. R. Chien, and D. A. Melton. Marked differences in differentiation propensity among human embryonic stem cell lines. *Nat Biotechnol*, 26(3):313–5, 2008.

[100] S. L. Paige, S. Thomas, C. L. Stoick-Cooper, H. Wang, L. Maves, R. Sandstrom, L. Pabon, H. Reinecke, G. Pratt, G. Keller, R. T. Moon, J. Stamatoyannopoulos, and C. E. Murry. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell*, 151(1):221–32, 2012.

[101] E. Paradis, J. Claude, and K. Strimmer. Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–90, 2004.

[102] X. Peng, J. Thierry-Mieg, D. Thierry-Mieg, A. Nishida, L. Pipes, M. Bozinoski, M. J. Thomas, S. Kelly, J. M. Weiss, M. Raveendran, D. Muzny, R. A. Gibbs, J. Rogers, G. P. Schroth, M. G. Katze, and C. E. Mason. Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (nhprtr). *Nucleic Acids Res*, 43(Database issue):D737–42, 2015.

[103] I. Piccini, J. Rao, G. Seebohm, and B. Greber. Human pluripotent stem cell-derived cardiomyocytes: Genome-wide expression profiling of long-term in vitro maturation in comparison to human heart tissue. *Genom Data*, 4:69–72, 2015.

[104] K. S. Pollard, S. R. Salama, N. Lambert, M. A. Lambot, S. Coppens, J. S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel, A. D. Kern, C. Dehay, H. Igel, Jr. Ares, M., P. Vanderhaeghen, and D. Haussler. An rna gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):167–72, 2006.

[105] S. Prabhakar, A. Visel, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, H. Morrison, D. R. Fitzpatrick, V. Afzal, L. A. Pennacchio, E. M. Rubin, and J. P. Noonan. Human-specific gain of function in a developmental enhancer. *Science*, 321(5894):1346–50, 2008.

[106] S. L. Prescott, R. Srinivasan, M. C. Marchetto, I. Grishina, I. Narvaiza, L. Selleri, F. H. Gage, T. Swigut, and J. Wysocka. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, 163(1):68–83, 2015.

[107] J. Qin, M. Chang, S. Wang, Z. Liu, W. Zhu, Y. Wang, F. Yan, J. Li, B. Zhang, G. Dou, J. Liu, X. Pei, and Y. Wang. Connexin 32-mediated cell-cell communication is essential for hepatic differentiation from human embryonic stem cells. *Sci Rep*, 6:37388, 2016.

[108] A. Rada-Iglesias, R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut, and J. Wysocka. Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell*, 11(5):633–48, 2012.

[109] P. Rana, B. Anson, S. Engle, and Y. Will. Characterization of human-induced pluripotent stem cell-derived cardiomyocytes: bioenergetics and utilization in safety screening. *Toxicol Sci*, 130(1):117–31, 2012.

[110] J. Rao, M. J. Pfeiffer, S. Frank, K. Adachi, I. Piccini, R. Quaranta, M. Arauzo-Bravo, J. Schwarz, D. Schade, S. Leidel, H. R. Scholer, G. Seebohm, and B. Greber. Stepwise clearance of repressive roadblocks drives cardiac induction in human escs. *Cell Stem Cell*, 18(3):341–53, 2016.

[111] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol*, 14(9):R95, 2013.

[112] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. Gc-content normalization for rna-seq data. *BMC Bioinformatics*, 12:480, 2011.

[113] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47, 2015.

[114] Consortium Roadmap Epigenomics, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[115] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010.

[116] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[117] M. V. Rockman, M. W. Hahn, N. Soranzo, F. Zimprich, D. B. Goldstein, and G. A. Wray. Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol*, 3(12):e387, 2005.

[118] R. Roessler, S. A. Smallwood, J. V. Veenvliet, P. Pechlivanoglou, S. P. Peng, K. Chakrabarty, M. J. Groot-Koerkamp, R. J. Pasterkamp, E. Wesseling, G. Kelsey, E. Boddeke, M. P. Smidt, and S. Copray. Detailed analysis of the genetic and epigenetic signatures of ipsc-derived mesodiencephalic dopaminergic neurons. *Stem Cell Reports*, 2(4):520–33, 2014.

[119] I. G. Romero, I. Ruvinsky, and Y. Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*, 13(7):505–16, 2012.

[120] J. Ruiz-Orera, J. Hernandez-Rodriguez, C. Chiva, E. Sabido, I. Kondova, R. Bontrop, T. Marques-Bonet, and M. M. Alba. Origins of de novo genes in human and chimpanzee. *PLoS Genet*, 11(12):e1005721, 2015.

[121] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom. Chip-seq: using high-throughput sequencing to discover protein-dna interactions. *Methods*, 48(3):240–8, 2009.

[122] M. D. Schultz, Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker. Human body epigenome maps reveal noncanonical dna methylation variation. *Nature*, 523(7559):212–6, 2015.

[123] K. B. Scotland, S. Chen, R. Sylvester, and L. J. Gudas. Analysis of rex1 (zfp42) function in embryonic stem cell differentiation. *Dev Dyn*, 238(8):1863–77, 2009.

[124] J. Shan, R. E. Schwartz, N. T. Ross, D. J. Logan, D. Thomas, S. A. Duncan, T. E. North, W. Goessling, A. E. Carpenter, and S. N. Bhatia. Identification of small molecules for human hepatocyte expansion and ips differentiation. *Nat Chem Biol*, 9(8):514–20, 2013.

[125] T. Shinozawa, A. Tsuji, K. Imahashi, K. Nakashima, H. Sawada, H. Toyoshiba, S. Yamamoto, K. Takami, and R. Imai. Gene expression profiling of functional murine embryonic stem cell-derived cardiomyocytes and comparison with adult heart: profiling of murine esc-derived cardiomyocytes. *J Biomol Screen*, 14(3):239–45, 2009.

[126] S. J. Sholtis and J. P. Noonan. Gene regulation and the origins of human biological uniqueness. *Trends Genet*, 26(3):110–8, 2010.

[127] M. L. Siegal and A. Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proc Natl Acad Sci U S A*, 99(16):10528–32, 2002.

[128] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[129] G. K. Smyth, J. Michaud, and H. S. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–75, 2005.

[130] M. Somel, H. Franz, Z. Yan, A. Lorenc, S. Guo, T. Giger, J. Kelso, B. Nickel, M. Dannemann, S. Bahn, M. J. Webster, C. S. Weickert, M. Lachmann, S. Paabo, and P. Khaitovich. Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A*, 106(14):5743–8, 2009.

[131] C. A. Sommer, C. Christodoulou, A. Gianotti-Sommer, S. S. Shen, B. S. Sailaja, H. Hezroni, A. Spira, E. Meshorer, D. N. Kotton, and G. Mostoslavsky. Residual expression of reprogramming factors affects the transcriptional program and epigenetic signatures of induced pluripotent stem cells. *PLoS One*, 7(12):e51711, 2012.

[132] M. Y. Son, H. Choi, Y. M. Han, and Y. S. Cho. Unveiling the critical role of rex1 in the regulation of human stem cell pluripotency. *Stem Cells*, 31(11):2374–87, 2013.

[133] C. Soneson and M. Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14(1):91, 2013.

[134] Q. Song and A. D. Smith. Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–1, 2011.

[135] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5, 2003.

[136] John D. Storey, Jonathan E. Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B (Methodological)*, 66(Part 1):187–205, 2004.

[137] K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–72, 2007.

[138] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76, 2006.

[139] A. K. Teo, S. J. Arnold, M. W. Trotter, S. Brown, L. T. Ang, Z. Chng, E. J. Robertson, N. R. Dunn, and L. Vallier. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev*, 25(3):238–50, 2011.

[140] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck. A beta-mixture quantile normalization method for correcting

probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–96, 2013.

[141] S. Tohyama, F. Hattori, M. Sano, T. Hishiki, Y. Nagahata, T. Matsuura, H. Hashimoto, T. Suzuki, H. Yamashita, Y. Satoh, T. Egashira, T. Seki, N. Muraoka, H. Yamakawa, Y. Ohgino, T. Tanaka, M. Yoichi, S. Yuasa, M. Murata, M. Suematsu, and K. Fukuda. Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. *Cell Stem Cell*, 12(1):127–37, 2013.

[142] I. Tomioka, T. Maeda, H. Shimada, K. Kawai, Y. Okada, H. Igarashi, R. Oiwa, T. Iwasaki, M. Aoki, T. Kimura, S. Shiozawa, H. Shinohara, H. Suemizu, E. Sasaki, and H. Okano. Generating induced pluripotent stem cells from common marmoset (callithrix jacchus) fetal liver cells using defined factors, including lin28. *Genes Cells*, 15(9):959–69, 2010.

[143] C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–11, 2009.

[144] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, and F. Ponten. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

[145] H. Uosaki, P. Cahan, D. I. Lee, S. Wang, M. Miyamoto, L. Fernandez, D. A. Kass, and C. Kwon. Transcriptional landscape of cardiomyocyte maturation. *Cell Rep*, 13(8):1705–16, 2015.

[146] H. Uosaki and Y. H. Taguchi. Comparative gene expression analysis of mouse and human cardiac maturation. *Genomics Proteomics Bioinformatics*, 14(4):207–15, 2016.

[147] C. W. van den Berg, S. Okawa, S. M. Chuva de Sousa Lopes, L. van Iperen, R. Passier, S. R. Braam, L. G. Tertoolen, A. del Sol, R. P. Davis, and C. L. Mummery. Transcriptome of human foetal heart compared with cardiomyocytes from pluripotent stem cells. *Development*, 142(18):3231–8, 2015.

[148] C. C. Veerman, G. Kosmidis, C. L. Mummery, S. Casini, A. O. Verkerk, and M. Bellin. Immaturity of human stem-cell-derived cardiomyocytes in culture: fatal flaw or soluble problem? *Stem Cells Dev*, 24(9):1035–52, 2015.

[149] M. Viotti, S. Nowotschin, and A. K. Hadjantonakis. Sox17 links gut endoderm morphogenesis and germ layer segregation. *Nat Cell Biol*, 16(12):1146–56, 2014.

[150] C. H. Waddington. Canalization of development and genetic assimilation of acquired characters. *Nature*, 183(4676):1654–5, 1959.

[151] CH Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565, 1942.

[152] J. A. Wamstad, J. M. Alexander, R. M. Truty, A. Shrikumar, F. Li, K. E. Eilertson, H. Ding, J. N. Wylie, A. R. Pico, J. A. Capra, G. Erwin, S. J. Kattman, G. M. Keller, D. Srivastava, S. S. Levine, K. S. Pollard, A. K. Holloway, L. A. Boyer, and B. G. Bruneau. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*, 151(1):206–20, 2012.

[153] G. Wang, M. L. McCain, L. Yang, A. He, F. S. Pasqualini, A. Agarwal, H. Yuan, D. Jiang, D. Zhang, L. Zangi, J. Geva, A. E. Roberts, Q. Ma, J. Ding, J. Chen, D. Z. Wang, K. Li, J. Wang, R. J. Wanders, W. Kulik, F. M. Vaz, M. A. Laflamme, C. E. Murry, K. R. Chien, R. I. Kelley, G. M. Church, K. K. Parker, and W. T. Pu. Modeling the mitochondrial cardiomyopathy of barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat Med*, 20(6):616–23, 2014.

[154] Y. Wang, J. Qin, S. Wang, W. Zhang, J. Duan, J. Zhang, X. Wang, F. Yan, M. Chang, X. Liu, B. Feng, J. Liu, and X. Pei. Conversion of human gastric epithelial cells to multipotent endodermal progenitors using defined small molecules. *Cell Stem Cell*, 19(4):449–461, 2016.

[155] Y. Wei, T. Tenzen, and H. Ji. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, 16(1):31–46, 2015.

[156] M. Wernig, M. Gotz, and K. Eto. Overcoming ipsc obstacles. *Cell Stem Cell*, 19(3):291–292, 2016.

[157] Hadley Wickham. *ggplot2*. Springer, New York, New York, 2009.

[158] Z. Williams, P. Morozov, A. Mihailovic, C. Lin, P. K. Puvvula, S. Juranek, Z. Rosenwaks, and T. Tuschl. Discovery and characterization of pirnas in the human fetal ovary. *Cell Rep*, 13(4):854–863, 2015.

[159] J. Wu, D. Okamura, M. Li, K. Suzuki, C. Luo, L. Ma, Y. He, Z. Li, C. Benner, I. Tamura, M. N. Krause, J. R. Nery, T. Du, Z. Zhang, T. Hishida, Y. Takahashi, E. Aizawa, N. Y. Kim, J. Lajara, P. Guillen, J. M. Campistol, C. R. Esteban, P. J. Ross, A. Saghatelian, B. Ren, J. R. Ecker, and J. C. Izpisua Belmonte. An alternative pluripotent state confers interspecies chimaeric competency. *Nature*, 521(7552):316–21, 2015.

[160] Y. Wu, A. Mishra, Z. Qiu, S. Farnsworth, S. D. Tardif, and P. J. Hornsby. Nonhuman primate induced pluripotent stem cells in regenerative medicine. *Stem Cells Int*, 2012:767195, 2012.

[161] Y. Wu, Y. Zhang, A. Mishra, S. D. Tardif, and P. J. Hornsby. Generation of induced pluripotent stem cells from newborn marmoset skin fibroblasts. *Stem Cell Res*, 4(3):180–8, 2010.

[162] S. Wunderlich, M. Kircher, B. Vieth, A. Haase, S. Merkert, J. Beier, G. Gohring, S. Glage, A. Schambach, E. C. Curnow, S. Paabo, U. Martin, and W. Enard. Primate ips cells as tools for evolutionary analyses. *Stem Cell Res*, 12(3):622–9, 2014.

[163] N. Xia, P. Zhang, F. Fang, Z. Wang, M. Rothstein, B. Angulo, R. Chiang, J. Taylor, and R. A. Reijo Pera. Transcriptional comparison of human induced and primary midbrain dopaminergic neurons. *Sci Rep*, 6:20270, 2016.

[164] W. Xie, M. D. Schultz, R. Lister, Z. Hou, N. Rajagopal, P. Ray, J. W. Whitaker, S. Tian, R. D. Hawkins, D. Leung, H. Yang, T. Wang, A. Y. Lee, S. A. Swanson, J. Zhang, Y. Zhu, A. Kim, J. R. Nery, M. A. Urich, S. Kuan, C. A. Yen, S. Klugman, P. Yu, K. Suknuntha, N. E. Propson, H. Chen, L. E. Edsall, U. Wagner, Y. Li, Z. Ye, A. Kulkarni, Z. Xuan, W. Y. Chung, N. C. Chi, J. E. Antosiewicz-Bourget, I. Slukvin, R. Stewart, M. Q. Zhang, W. Wang, J. A. Thomson, J. R. Ecker, and B. Ren. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–48, 2013.

[165] A. G. Xu, L. He, Z. Li, Y. Xu, M. Li, X. Fu, Z. Yan, Y. Yuan, C. Menzel, N. Li, M. Somel, H. Hu, W. Chen, S. Paabo, and P. Khaitovich. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by rna-seq. *PLoS Comput Biol*, 6:e1000843, 2010.

[166] Gilad Y and Mizrahi-Man O. A reanalysis of mouse encode comparative gene expression data. *F1000 Research*, 4(121):doi: 10.12688/f1000research.6536.1, 2015.

[167] L. Yan, H. Guo, B. Hu, R. Li, J. Yong, Y. Zhao, X. Zhi, X. Fan, F. Guo, X. Wang, W. Wang, Y. Wei, Y. Wang, L. Wen, J. Qiao, and F. Tang. Epigenomic landscape of human fetal brain, heart, and liver. *J Biol Chem*, 291(9):4386–98, 2016.

[168] X. Yang, M. Rodriguez, L. Pabon, K. A. Fischer, H. Reinecke, M. Regnier, N. J. Sniadecki, H. Ruohola-Baker, and C. E. Murry. Tri-iodo-l-thyronine promotes the maturation of human cardiomyocytes-derived from induced pluripotent stem cells. *J Mol Cell Cardiol*, 72:296–304, 2014.

[169] A. A. Youssef, E. G. Ross, R. Bolli, C. J. Pepine, N. J. Leeper, and P. C. Yang. The promise and challenge of induced pluripotent stem cells for cardiovascular applications. *JACC Basic Transl Sci*, 1(6):510–523, 2016.

[170] G. Yu, L. G. Wang, Y. Han, and Q. Y. He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS*, 16(5):284–7, 2012.

[171] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

[172] D. Zhao, S. Chen, S. Duo, C. Xiang, J. Jia, M. Guo, W. Lai, S. Lu, and H. Deng. Promotion of the efficient metabolic maturation of human pluripotent stem cell-derived hepatocytes by correcting specification defects. *Cell Res*, 23(1):157–61, 2013.

[173] P. Zhou and W. T. Pu. Recounting cardiac cellular composition. *Circ Res*, 118(3):368–70, 2016.

[174] X. Zhou, C. E. Cain, M. Myrthil, N. Lewellen, K. Michelini, E. R. Davenport, M. Stephens, J. K. Pritchard, and Y. Gilad. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol*, 15(12):547, 2014.

[175] M. J. Ziller, F. Muller, J. Liao, Y. Zhang, H. Gu, C. Bock, P. Boyle, C. B. Epstein, B. E. Bernstein, T. Lengauer, A. Gnirke, and A. Meissner. Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. *PLoS Genet*, 7(12):e1002389, 2011.

[176] A. M. Zorn and J. M. Wells. Vertebrate endoderm development and organ formation. *Annu Rev Cell Dev Biol*, 25:221–51, 2009.