THE UNIVERSITY OF CHICAGO

USING FUNCTIONAL GENOMICS APPROACHES TO INVESTIGATE HUMAN COMPLEX TRAITS AND DISEASES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
LAUREN ELIZABETH BLAKE

CHICAGO, ILLINOIS
MARCH 2020

Table of Contents

List of Figures

List of Tables [1]

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

ACKNOWLEDGMENTS

They say it takes a village to raise a child. In much the same way, it takes a village to get a PhD. I am extremely grateful for my "village" of supporters, including my mentors, collaborators, friends, and family.

I am very thankful to my advisor, Yoav Gilad. He has been extremely influential on me as a scientist, teaching me how to creatively solve challenging problems in human genetics with rigor and focus. His support along the path to graduation has been critical.

I am thankful to my committee members, Oni Basu, Marcelo Nobrega, and Matthew Stephens. I am also grateful for the opportunity to work with top eating disorder specialists, including Jennifer Wildes, Jessica Baker, and Cindy Bulik. This work would not have been possible without the Human Genetics community in Cummings Life Science Center. Additionally, I have enjoyed working with with Chris Porras, Dan Rice, Eric Friedlander, and Roy Morgan on the Computational STEM Lab (CSL) outreach program.

Throughout my PhD, I have been financially supported by the Genetics and Regulation Training Grant (T32GM007197) from the National Institutes of Health, the National Science Foundation (Graduate Research Fellowship DGE-1144082 and Innovation Corps NSF Award 1735858), and the University of North Carolina Center of Excellence for Eating Disorders. John Novembre also supported the CSL. Sue Levison started as my "grants person" and evolved into a friend. Sue made sure that I didn't run away during my first week of graduate school, and has seen me all the way through graduation.

After about 2 weeks of my rotation in the Gilad lab, I knew that this was where I wanted to do my thesis research. After 5 years, I am still confident that I made the right decision. This feeling is in large part due to my fantastic labmates, past and present. In particular, thank you to Nick Banovich, Seb Pott, Genevieve Housman, Kenneth Barr, Benjamin Fair, Reem Elorbany, Katie Rhodes, Sidney Wang, Po-Young Tung, John Blischak, Ittai Eres, Natalia Gonzales, Jonathan Burnett, Emily Davenport, and Samantha Thomas. Michelle

Ward has influenced my scientific thinking tremendously. Briana Mittleman was a fantastic baymate, and I will miss the Duke corner of lab tremendously. I am also thankful to the technicians in the lab, particularly Amy Mitrano, Marsha Myrthil, and Claudia Chavarria, who also helped with my thesis work. I'd also like to thank my close collaborators, particularly Julien Roux and Bryan Pavlovic, without whom Chapters 2 and 3 in this thesis would not have been possible. Joyce Hsiao and Abhishek Sarkar also provided key statistical advice for my work.

Thank you to my friends Unjin Lee, Charlie Lang, Marissa Pelot, Hillary Childs, Sierra Smart, Sam Lohnes, Tatiana Birgisson, Robin Schwartzman, and Janet Scognamiglio. Marcus Soliai and Aarti Venkat are best coffee buddies anyone could ask for. It is an honor to call Manny Vazquez my friend and have gotten to collaborate with him. Finally, it would be impossible to list all the ways that Bryce van de Geijn supported me during my Phd.

My family has been incredibly supportive. In particular, my grandparents, Nannie and Papa, encouraged me via weekly phone calls throughout my Phd. Emily and Kate Blake are both amazing people, and I am lucky to call them my sisters. Thank you to my Mom and Dad, who have loved and encouraged me during my entire life.

ABSTRACT

A primary aim of human genetics is to determine how genetic variation impacts phenotypic variation, including in complex traits and diseases. Understanding this relationship will ultimately allow us to understand the molecular basis of complex traits and better diagnose and treat human diseases. To dissect this relationship, human geneticists have leveraged comparisons between humans and other primates, as well as between different groups of humans. To maximize the utility of functional genomics studies within and between species, proper study design is essential. In Chapters 2 and 3, I describe two primate comparative studies that highlight a variety of study design challenges and discuss potential solutions.These studies demonstrate that adherence to key study design principles can facilitate biological insight. In Chapter 4, I apply these principles to a new problem, distinguishing individuals with eating disorders at high risk of rehospitalization from those with lower risk. In the final chapter, I discuss lessons learned and suggest the next steps for using functional genomics to study eating disorders.

CHAPTER 1

INTRODUCTION

1.1 Using genetics to understand complex traits and diseases

The overarching goal of human genetics is to understand how genetic variation influences phenotypes– including complex traits and diseases– within our species. One approach to connect genotypic to phenotypic variation is to compare humans with non-human primates [118]. This approach is particularly compelling given that human share over 99% of their DNA sequence with chimpanzees [187, 157]. Efforts have been made to understand the 1%, that is to say the impact of these approximately 30 million single nucleotide polymorphisms (SNPs) differences [9, 187, 157, 167]. Importantly, the majority of these SNPs are in the non-protein coding regions of the genome [118]. Overall this high degree of sequence similarity led Mary Claire King and Greg Wilson to hypothesize that it is the way these DNA sequences are regulated that leads to phenotypic differences [118]. Consequently, comparisons between humans and non-human primates can help to determine how these DNA sequences function and to reveal the underlying mechanisms that act on or as a result of sequence variation. This desire for for mechanistic understanding has inspired a number of studies to compare gene regulation levels, epigenetic marks, protein levels, and various other molecular phenotypes [14, 24, 159, 188, 187, 40, 91, 92, 93, 153, 189, 64, 213, 226]. These efforts have resulted in large comparative genomic catalogs of similarities and differences in gene regulation between humans and other primates. These catalogs have great potential to help us better understand the evolutionary processes that led to adaptations in humans [9, 23, 24, 39, 130, 188, 84, 108, 187, 112, 115, 122, 151, 153, 177], establish informed models of the relative importance of changes in different molecular mechanisms to regulatory evolution [113, 214], and identify molecular pathways that may be functionally important in the context of complex diseases [23, 24]. However, the genetic variants that drive phenotypic

diversity often have indirect effects on gene expression or protein levels [118]. It is therefore essential to design comparative studies that allow us to isolate the variables of interest while minimizing the effects of unwanted biological and technical differences. Yet, all too often, various aspects of study design are overlooked, to the detriment of the field. In extreme cases, poor study design leads to erroneous inference and incorrect interpretations [124, 220]. The majority of the time, a poor study design limits the accuracy of the study and by extension, the biological insight that can be drawn from it [23, 40, 155]. Compounding the issue is that study design considerations are usually not explicitly discussed in comparative genomic papers.

The majority of my work has focused on primate comparative genomics - an area which presents many opportunities to discuss how an effective study design can affect the results of a study and aid in its interpretation. In this Introduction, I focus on two principles critical to identifying robust biological differences between species: minimizing confounders and careful sample collection. However, the principles of study design that I will discuss extend beyond primate comparative genomics to any study that makes comparisons between groups, including case-control studies in humans.

1.2 Study design challenges in comparative primate genomics

Confounders and other potential biases. As the technology used in genomic studies has progressed, so too has our understanding of the widespread nature and impact of confounders and other potential biases. Furthermore, using multiple species in functional genomic studies has increased the difficulty of minimizing these confounders. For example, early comparative studies using gene expression and DNA methylation microarrays often did not account for the attenuation of hybridization caused by sequence mismatches, which differ between species [39, 188, 76, 112]. More recently, common confounders of sequence-based comparative studies include individual sampling schemes that are unbalanced across species, and

sample processing steps that are segregated by species [23, 159, 40, 47, 124, 224]. Systemic differences inherent to the samples, such as differences in material quality between species [159, 155], also remain a concern. Similarly, a primate comparative framework brings forth analytical challenges. For example, analyses that do not use orthologous sequences or effective normalization procedures can result in bias [23, 40, 155]. Yet, most comparative genomic studies of humans and non-human primates that we are aware of, including previous studies from our own group [23, 40, 155], suffer from one or more of these weaknesses and caveats. Very few people would disagree that it is important to use good study design. However, the number of potential confounders are vast, including sex, date of death, age, RNA concentration, RIN score, RNA extraction date, library concentration, index sequence, sequencing pool, sequencing location, sequencing lane, total sequencing reads. Therefore, it can be difficult to detect confounding factors and bias introduced during sample processing or data analysis. Unfortunately, these factors are often neither accounted for nor discussed. Sometimes, this can lead to erroneous conclusions, which could have major implications for biological research. For example, a recent paper claimed that global gene expression levels were driven by species rather than tissue type [124]. Upon re-analysis, it was uncovered that the human and mouse samples used in the study were sequenced in different batches [220].With this study design, the biological variable of interest (species) is confounded with the technical variable of sequence batch. Therefore, it is unknown to what extent the technical variable drove the biological results reported in the original paper. Sorting out which results were driven by biological, rather than technical differences, is often led to the reader. Such a task can be quite challenging, as the comparative genomics field, and indeed, the larger genomics community, lacks consensus regarding meta-data collection and study documentation, particularly around sample and study design reporting.

Opportunistic study collection. Sample type, size, and collection techniques are critical study design considerations in comparative studies of primates. Until recently, flash frozen

tissues were one of the only options for comparing primate biological material [81]. Because of the difficulty of obtaining samples (both for logistic and ethical reasons), we could typically sample only a small number of individuals from each species. Consequently, some primate comparative studies only have only a handful of individuals per species [24, 92, 153]. Particularly problematic is when there is only 1 individual per species, as individual and tissue are confounded. Even when there are 3 or more individuals per species, tissues are often subject to high environmental variances because the donors are in an uncontrolled environment, and also flash frozen and shipped post mortem [167]. The necessity of collecting samples opportunistically, together with small sample sizes, can lead to incomplete power to detect regulatory differences between species in any given study, and hence to relatively large apparent differences between studies. Furthermore, it was nearly impossible to obtain multiple tissue samples from the same individual. For example, to date, there have been no published comparative studies in primates that have analyzed multiple tissues sampled from the same individuals across multiple species in a balanced design [167]. Consequently, regulatory differences between tissues are always confounded with regulatory differences between individuals [153]. In turn, relative measures of tissue-specific regulatory differences between species are confounded with inter-tissue differences in regulatory variation within species.

1.3 Study design challenges in psychiatric genetics

Unfortunately, these study design challenges are not limited to the field of primate comparative genomics. For example, these are issues in psychiatric genetics and are highly prevalent in the nascent field of eating disorders genomics.

Confounders and other potential biases. When performing eating disorder studies, complications with the phenotype lend itself to confounders and other potential biases. Like other psychiatric disorders, the way that eating disorders are diagnosed is inherently subjective, even if clinicians use the same criteria [22, 143]. For example, AN diagnoses comprise of qual-

itative criteria, such as over-concern with shape and weight and fear of becoming overweight [5]. While there are attempts to measure these aspects, it is not difficult. Consequently, there was an emphasis on introducing more quantitative measures, such as body mass index (BMI) to the diagnosis [54]. However, these types of quantitative measures are also inherently tricky for psychiatric disorders. For example, in AN, it is unknown whether BMI is only a consequence of AN or plays a role in furthering the disorder [54]. In case-control studies of individuals with anorexia nervosa (AN), disease state is confounded with BMI [62, 215]. This confounding complicates analysis and interpretation of results. For example, when a SNP is associated with AN cases, is the SNP associated with a specific AN behavior (and if so, which one) or with BMI? Even within cases, clinical variables can be confounded, including medication types, age of diagnosis, and number of rehospitalizations [117]. In addition to the large number of biological variables, there are also technical considerations, such as RIN score and batch. As discussed earlier, the larger genomics community lacks consensus regarding meta-data collection and study documentation. This area has also been discussed in the eating disorder field but have not been widely acted upon [103].

Opportunistic study collection. Similar to primates, tissues from humans are collected opportunistically. For many complex psychiatric trait, it is difficult to link genetic variation or gene regulatory differences, to phenotypes. While gene regulation in the brain is a logical place to start, it is difficult to access brains from living patients and a brain bank for eating disorders has only recently been established [58]. Moreover, the brain is extremely heterogeneous. Consequently, it is difficult to identify which brain regions are the most relevant to study [221]. Moreover, while recent research suggests tissues besides the brain may be relevant to disease state in AN [63, 102, 215], it is hard to decide which tissue to study. Blood is relatively easy to collect from individuals with eating disorders, but the clinical utility of whole blood is unknown [117]. Since individuals are collected opportunistically, collecting large sample sizes, particularly in longitudinal studies are difficult. Furthermore, individuals

are almost always recruited during a state of illness, so it can be difficult to predict patient trajectory and therefore the outcomes represented in the cohort [75]. This issue is compounded by the fact that many eating-disorder phenotypes, including treatment outcomes, are nebulous [114].Indeed, although the field of eating disorder genomics faces many of the same obstacles as primate comparative genomics and psychiatric genomics more generally, it also presents a unique set of challenges.

1.4 Conclusion

Much of my thesis work has been devoted to addressing these challenges of potential biases and opportunistic study collection. In Chapters 2 and 3, I present examples of how to approach these challenges in gene regulatory studies across primates. Furthermore, I demonstrate that adherence to these key study design principles helps elucidate biological insight. In Chapter 4, I apply these lessons learned in a new context: comparing individuals with eating disorders at high risk of rehospitalization versus those with lower risk. In the final chapter, I discuss opportunities and challenges when using functional genomics to study eating disorders.

CHAPTER 2

A COMPARISON OF GENE EXPRESSION AND DNA METHYLATION PATTERNS ACROSS TISSUES AND SPECIES

2.1 Abstract[1]

Previously published comparative functional genomic data sets from primates using frozen tissue samples, including many data sets from our own group, were often collected and analyzed using non-optimal study designs and analysis approaches. In addition, when samples from multiple tissues were studied in a comparative framework, individual and tissue were confounded. We designed a multi-tissue comparative study of gene expression and DNA methylation in primates that minimizes confounding effects, by using a balanced design with respect to species, tissues, and individuals. We also developed a comparative analysis pipeline that minimizes biases due to sequence divergence. Thus, we present the most comprehensive catalog of similarities and differences in gene expression and methylation levels between livers, kidneys, hearts, and lungs, in humans, chimpanzees, and rhesus macaques. We estimate that overall, only between 7 to 11% (depending on the tissue) of inter-species differences in gene expression levels can be accounted for by corresponding differences in promoter DNA methylation. However, gene expression divergence in conserved tissue-specific genes can be explained by corresponding inter-species methylation changes more often. Finally, we show that genes whose tissue-specific regulatory patterns are consistent with the action of natural selection are highly connected in both gene regulatory and protein-protein interaction networks.

1. Citation for chapter: Blake LE*, Roux J*, Hernando-Herraez I, Banovich NE, Garcia-Perez R, Hsiao CJ, Eres I, Chavarria C, Marques-Bonet T, Gilad Y. A comparison of gene expression and DNA methylation patterns across tissues and species. Accepted, *Genome Research.* * denotes equal contribution.

2.2 Introduction

Gene regulatory differences between humans and other primates are hypothesized to underlie human-specific traits [118]. Over the past decade, dozens of comparative genomic studies focused on characterizing mRNA expression level differences between primates in a large number of tissues (e.g., [14, 24, 159, 188, 111]), typically focusing on differences between humans and other primates. A few studies have also characterized inter-primate differences in regulatory mechanisms and phenotypes other than gene expression levels, such as DNA methylation levels, chromatin modifications and accessibility, and protein expression levels [40, 91, 92, 93, 153, 189, 208, 213, 226]. These studies often construct catalogs of gene expression levels and other mechanisms. These catalogs have been useful to better understand the evolutionary processes that led to adaptations in humans [9, 23, 24, 39, 42, 188, 84, 108, 111, 112, 115, 122, 151, 153, 177] and ancestral or derived phenotypes that may be relevant to human diseases [127, 167]. One caveat that is shared among practically all comparative studies in primates is related to difficulty in obtaining multiple tissue samples from the same individual. To date, there have been no published comparative studies in primates that have analyzed multiple tissues sampled from the same individuals across multiple species in a balanced design [167]. As a result, regulatory differences between tissues are always confounded with regulatory differences between individuals [24, 159, 46, 153]. In turn, catalogs from these studies can not be used to compare tissue-specific regulatory differences between species to inter-tissue differences in regulatory variation within species (see Discussion in [153]). Our group and others often use previously published catalogs of comparative data in primates in our different studies. While we do not expect previously observed patterns to be erroneous, we are aware that data on gene-specific inter-species regulatory differences, and especially data that pertain to comparisons of divergence across tissues, may be inaccurate for the reasons we discussed above. We thus designed the current study to produce a new comprehensive catalog of comparative gene

expression and DNA methylation data from humans, chimpanzees, and rhesus macaques, attempting to minimize possible confounders. The goal of our study is not to challenge previous conclusions or document specific differences between the current and previous data. Instead, we aim to provide a new and more accurate comparative catalog of inter-tissue and inter-species differences in gene regulation between humans and other primates, with substantial sample and study design documentation. Overall, we believe that this catalog can be useful for many future applications and can serve as a new benchmark for regulatory divergence in primates.

2.3 Results

2.3.1 Study design and data collection

To comparatively study gene expression levels and DNA methylation patterns in primates, we collected primary heart, kidney, liver and lung tissue samples from four human, four chimpanzee, and four rhesus macaque individuals (Figure 2.1A, Table 2.2A). From these 48 samples, we harvested RNA and DNA in parallel (Methods). After confirming that the RNA from all samples was of acceptable quality (Table 2.2B; Figure 2.5), we performed RNA-sequencing to obtain estimates of gene expression levels. Additional details about the donors, tissue samples, sample processing, and sequencing information can be found in the Methods and Table 2.2. We estimated gene expression levels using an approach designed to prevent biases driven by sequence divergence across the species (similar to the approach of [21]). Briefly, we first mapped RNA-sequencing reads to each species' respective genome, and to compare gene expression levels across species, we only calculated the number of reads mapping to exons that can be classified as clear orthologs across all three species (Table 2.2B). We excluded data from genes that were lowly expressed in over half of the samples as well as data from one human heart sample that was an obvious outlier, probably due to a

sample swap (Figure 2.6). We normalized the distribution of gene expression levels to remove systematic expression differences between species (maximizing the number of genes with invariant expression levels across species corresponds to our null hypothesis; see Methods). Through this process, we obtained TMM- and cyclic loess-normalized log2 counts per million (CPM) values for 12,184 orthologous genes to be used in downstream analyses (Table 2.3). Elements of study design, including sample processing, have previously been shown to impact gene expression data [220]. Consequently, we tested the relationship between a large number of technical factors recorded throughout our experiments and the biological variables of interest in our study, namely tissue and species (Methods, Supplemental Materials, Table 2.4A-B). We found that there were no technical confounders with tissue but two technical factors were confounded with species: time postmortem until collection and RNA extraction date (Figure 2.5B-C). Due to the opportunistic nature of sample collection, these confounders are practically impossible to avoid in comparative studies in primates (especially apes). We discuss possible implications of these confounders throughout the paper.

2.3.2 Gene expression varies more across tissues than across species

We first examined broad patterns in the gene expression data. A principal component analysis (PCA) indicated that, as expected [15, 159, 140], the primary sources of gene expression variation are tissue (Figure 2.1B, regression of PC1 by tissue = 0.81; P $<10^{-14}$; regression of PC2 by tissue = 0.70; P $<10^{-10}$; Tables 2.2A-B and 2.4A-B), followed by species (regression of PC2 by species = 0.27; P $<10^{-3}$; Tables 2.2A-B and 2.4A-B). This pattern is also supported by a clustering analysis based on the correlation matrix of pairwise gene expression estimates across samples (Figure 2.7). We then confirmed that, globally, gene expression levels across tissues from the same individual are more highly correlated than gene expression levels across tissues from different individuals (Figure 2.6C). This observation supports the intuitive notion that collecting and analyzing multiple tissues from the same individual

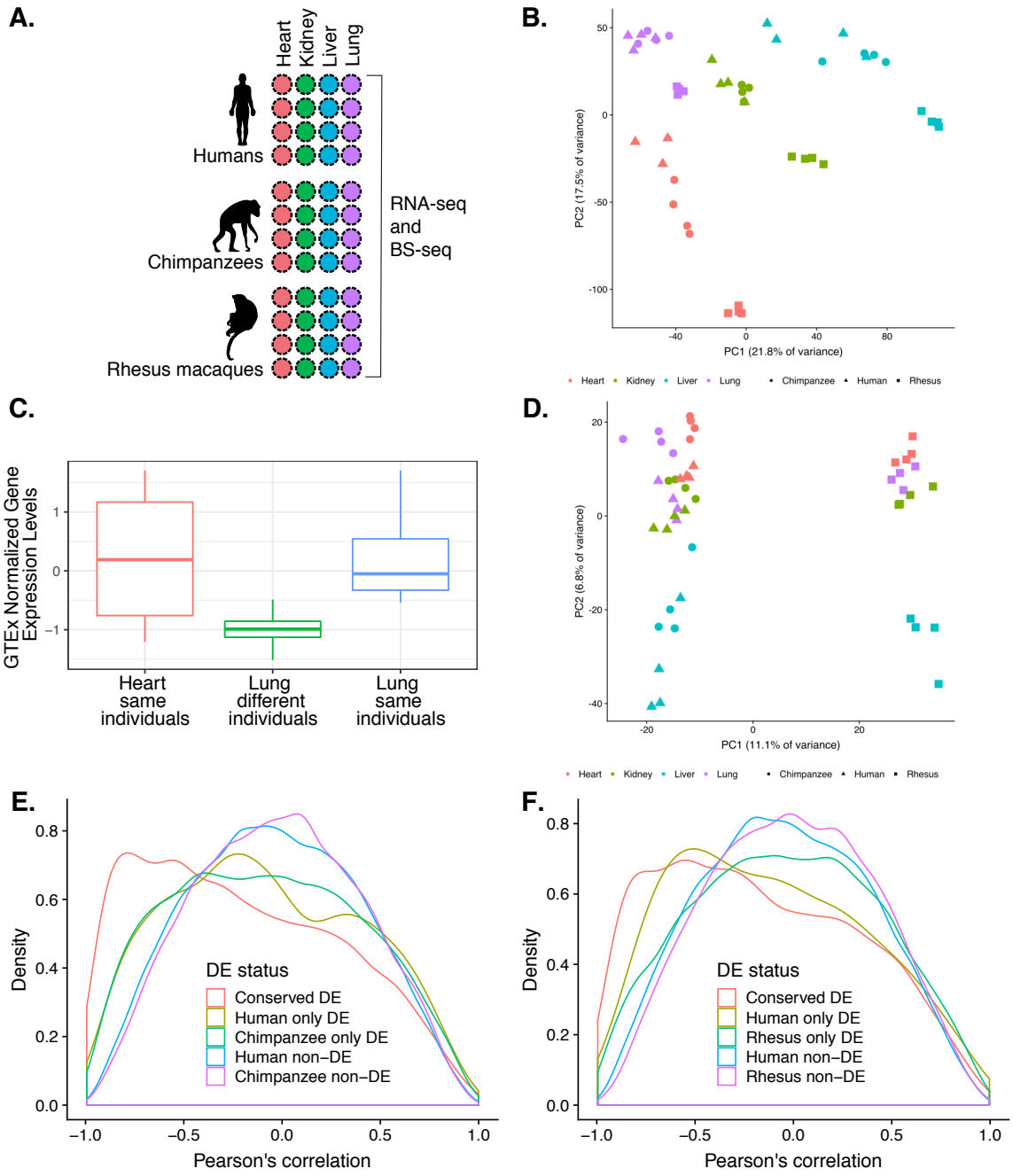


Figure 2.1: **Surveying gene expression and DNA methylation in diverse tissues across primates.** (A) Study design. (B) Principal components analysis (PCA) of gene expression levels in 47 samples. (C) Normalized gene expression from 4 donors in the GTEx heart collection compared to the lungs from different and the same donors' lungs in AC020922.1. (D) PCA of average methylation levels 250 bp upstream and downstream in 47 samples. (E-F.) Density function of the correlation between gene expression and DNA methylation levels in (E) human-chimpanzee and (F) humans and rhesus macaques orthologous genes.

is highly desirable in functional genomics studies. We sought further explicit evidence that a study design incorporated balanced collection of multiple tissues from the same individuals is more effective. To do so, we used data from the GTEx Consortium [87] for lung and heart. We first identified differentially expressed (DE) genes between lung and heart using all of the available GTEx data; we designated these classifications, which are based on hundreds of samples, as the "truth" (Supplemental Materials; Table 2.4E). Next, we repeatedly identified DE genes between lung and heart using GTEx data from randomly chosen sets of just 4 samples from each tissue, and compared the results to DE genes identified from an equivalent analysis of sets of 4 samples from each tissue, in which the tissue samples originated from the same donor. Compared to the "true classification" based on the entire GTEx dataset, DE analyses using data from samples of tissues that are matched for donors result in a higher ratio of true positives to false positives than analyses using samples from tissues that are unmatched for donors (P = 0.03; Table 2.4F). Given the small number of false positives in both datasets, study design is unlikely to impact large-scale, highly robust trends across species. However, this study design choice is particularly important if one is interested in individual genes (as demonstrated by an example in Figure 2.1C).

2.3.3 Putatively functional tissue-specific gene expression patterns

To analyze the pairwise regulatory differences across tissues and species, we used the framework of a linear model (see Methods). We first identified (at FDR = 1%) 3,695 to 7,027 (depending on the comparison we considered) differences in gene expression levels between tissues, within each species (Table 2.1A; Table 2.5). Overall, the patterns of inter-tissue differences in gene expression levels are similar in the three species, significantly more so than expected by chance alone ($P < 10^{-16}$, hypergeometric distribution; Supplemental Materials; Table 2.6). A range of 17 to 26% of inter-tissue DE genes have conserved inter-tissue expression patterns in all three species (Table 2.6). Regardless of species, we found the fewest

inter-tissue DE genes when we considered the contrast between liver and kidney, and the largest number of DE genes between liver and either heart or lung (Table 2.1A; Table 2.5B). Unfortunately, since our data were produced from bulk RNA-sequencing, we were unable to determine the impact of cell composition on the number of inter-tissue DE genes. We used the same framework of linear modeling to identify gene expression differences between species, within each tissue (Table 2.5A). Depending on the tissue and species we considered, we identified between 805 to 4,098 inter-species DE genes (at FDR = 1%; Table 2.1A, Table 2.5C). As expected given the known phylogeny of the three species, within each tissue, we classified far fewer DE genes between humans and chimpanzees than between either of these species and rhesus macaques (Table 2.5B). It is a common notion that genes with tissue-specific expression patterns may underlie tissue-specific functions. Previous catalogs of such patterns in primates were always confounded by the effect of individual variation (because each tissue was sampled from a different individual). To classify tissue-specific genes using our data, we focused on genes that are either up-regulated or down-regulated in a single tissue relative to the other three tissues (within one or more species). We define such genes as having a "tissue-specific" expression pattern, acknowledging that this definition may only be relevant in the context of the four tissues we considered here. Using this approach and considering the human data across all tissue comparisons, we identified 5,284 genes with tissue-specific expression patterns (FDR 1%, Figure 2.2). By performing similar analyses using the chimpanzee and rhesus macaque data, we found that the degree of conservation of tissue-specific expression patterns is higher than expected by chance ($P < 10^{-16}$; Figure 2.2). This observation is robust with respect to the statistical cutoffs we used to classify tissue-specific expression patterns (Table 2.7), indicating that many of these conserved tissue-specific regulatory patterns are likely of functional significance. To broadly analyze the biological function of genes with conserved tissue-specific expression, we performed a Gene Ontology enrichment analysis (GO, see Supplemental Materials). We found

these genes are indeed highly enriched with functional annotations that are relevant to the corresponding tissue (Table 2.8A-D, 2.9). For example, genes with conserved heart-specific expression patterns were enriched in GO categories related to muscle filament sliding (e.g. *ACTA1*, *MYL2*) and cardiac muscle contraction (e.g. *MYBPC3*, *TNNI3*).

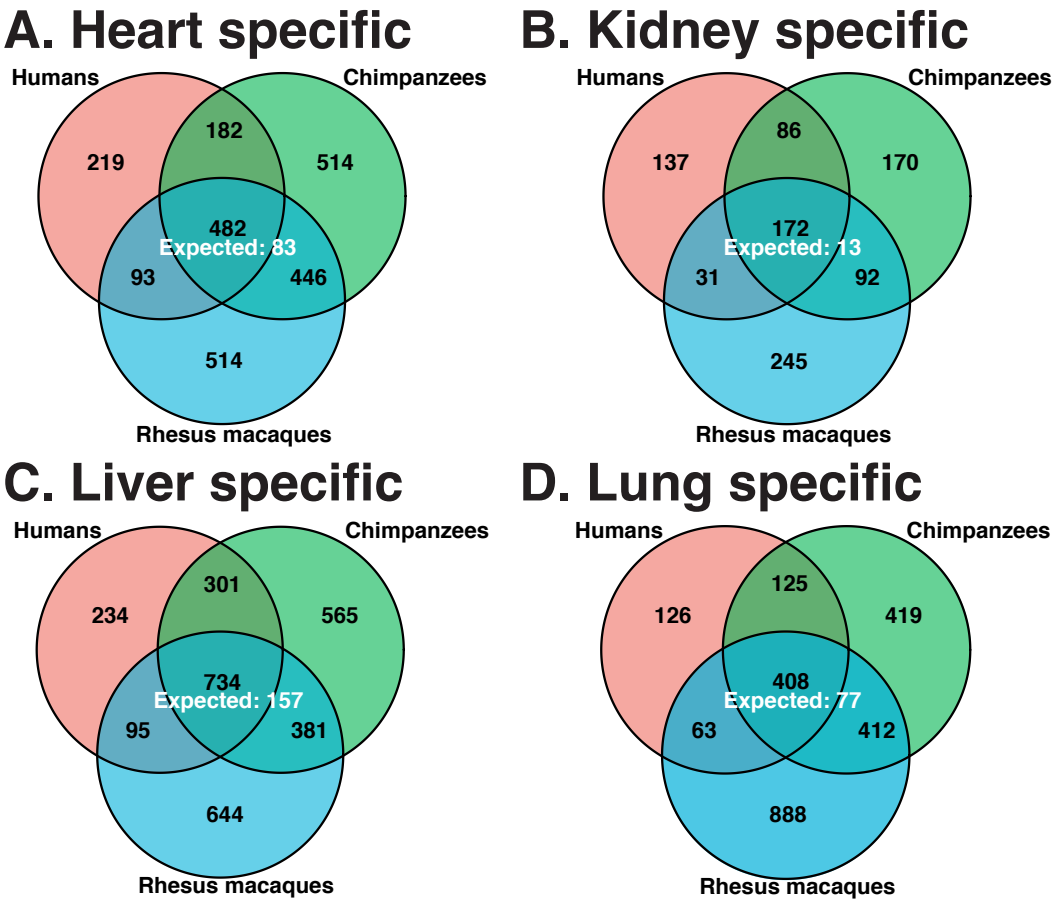


Figure 2.2: **The number of conserved tissue-specific DE genes across all three species, in the (A) heart, (B) kidney, (C) liver, and (D) lung, is greater than the number expected by chance.**

2.3.4 Functional Analysis of Gene Regulatory Differences

We sought further evidence that the classification of genes with conserved tissue specific expression patterns is meaningful. To do so, we considered transcription co-expression networks [194, 225] based on GTEx data from heart and lung [156]. We found that genes with conserved tissue specific expression patterns are more likely to appear as nodes in the networks than genes without tissue-specific expression patterns, or genes whose tissue-specific expression patterns are not conserved ($P < 10^{-5}$). When we only considered genes that do appear as nodes in the network, we found that genes with conserved tissue specific expression patterns are more likely to be classified as hubs in the networks than genes without tissue-specific expression patterns, or genes whose tissue-specific expression patterns is not conserved ($P < 0.007$). Motivated by these findings, we focused on gene expression patterns that are consistent with the action of natural selection (as described in [24]; see Supplemental Materials and Table 2.8E). We found that genes whose expression patterns are consistent with the action of either stabilizing or directional selection (top 10%; Table 2.8F) have more interactions with other genes in the network than genes whose expression patterns are not consistent with the action of natural selection (bottom 10%; $P < 0.05$ for all comparisons; Figure 2.2E). This observation is fairly robust with respect to percentile cutoff (Table 2.8F). We repeated a similar analysis by using protein-protein interaction data from the Human Protein Atlas [125, 201, 204, 223] in all four tissues. We again found that genes whose expression patterns are consistent with selection have more annotated protein-protein interactions ($P < 0.05$ in all 8 comparisons, Figure 2.2F; Table 8G). These interaction results suggest that functionally important genes are carefully regulated. Furthermore, this tight regulation occurs at both the gene expression and protein levels in primates.

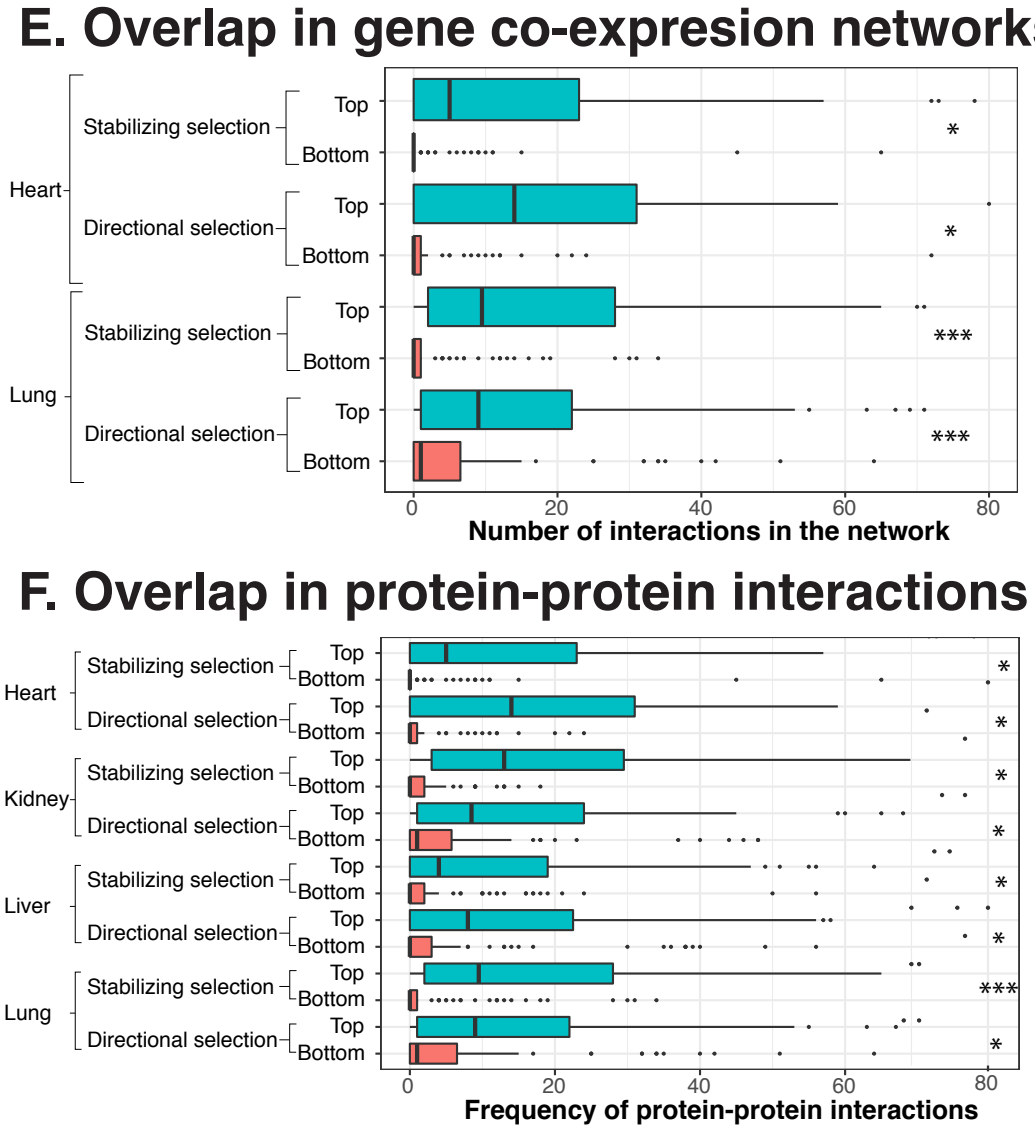


Figure 2.2, continued. In each tissue, genes with tissue-specific regulatory patterns that are consistent with the action of natural selection ("top") are more likely to appear in (E) gene co-expression networks and (F) have an increased number of protein-protein interactions those that are less consistent with the action of natural selection ("bottom"). * P <0.05, *** P <0.001. Note: The x-axes of 2E and 2F are cut off at 80 interactions for readability. In both cases, <5% of the data points are beyond this cutoff.

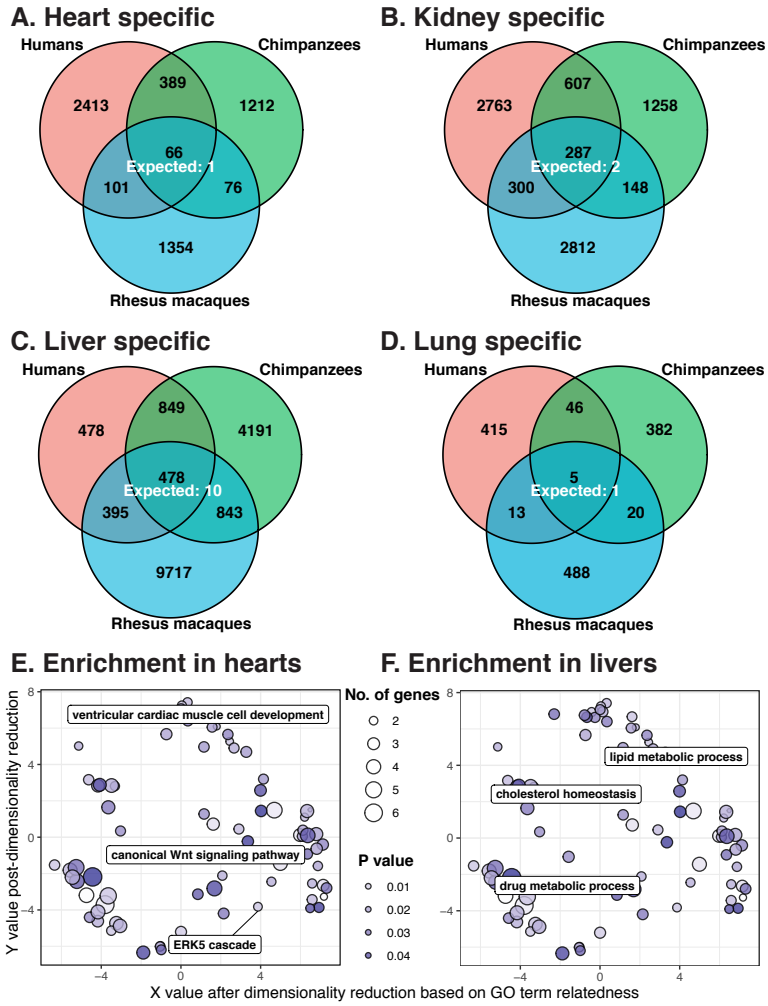


Figure 2.3: **Tissue-specific DMRs (FDR = 0.01).** The number of conserved tissue-specific DMRs in the (A) heart, (B) kidney, (C) liver, and (D) lung is greater than expected by chance. Genes with the closest TSSs to conserved tissue-specific DMRs are enriched for relevant functional annotations in (E) hearts and (F) livers.

Table 2.1A

DE between tissues (within species)	**Heart-Kidney**	**Heart-Liver**	**Heart-Lung**	**Kidney-Liver**	**Kidney-Lung**	**Liver-Lung**
Human	4224	4776	4037	4248	3695	4701
Chimpanzee	4971	6295	5365	4625	4623	6247
Rhesus macaque	5980	6933	6814	5126	5721	7027
DE between species	**Heart**	**Kidney**	**Liver**	**Lung**		
Human vs. Chimpanzee	2195	799	1364	805		
Human vs. Rhesus Macaque	4098	2347	2868	2833		
Chimpanzee vs. Rhesus Macaque	2781	2286	3139	1917		

Table 2.1B

Tissue DMRs (within species)	**Heart-Kidney**	**Heart-Liver**	**Heart-Lung**	**Kidney-Liver**	**Kidney-Lung**	**Liver-Lung**
Human	30291 (1.7%)	22561 (1.3%)	14208 (0.8%)	17910 (1.1%)	16521 (1.0%)	12842 (0.8%)
Chimpanzee	17699 (1.1%)	28767 (1.7%)	8524 (0.5%)	23847 (1.4%)	12076 (0.7%)	22107 (1.3%)
Rhesus macaque	23023 (1.5%)	41280 (2.7%)	7026 (0.5%)	35910 (2.4%)	15889 (1.1%)	32636 (2.1%)
Species DMRs	**Heart**	**Kidney**	**Liver**	**Lung**		
Human vs. Chimpanzee	14504 (1.1%)	10667 (0.8%)	9476 (0.8%)	8617 (0.7%)		
Human vs. Rhesus Macaque	25539 (5.3%)	21292 (4.5%)	21639 (4.5%)	17696 (3.9%)		

Table 2.1: **Pairwise DE genes and DMRs.** (A) Pairwise differentially expressed (DE) genes at FDR 1%. (B) Pairwise differentially methylated regions (DMRs) in autosomal chromosomes, percentage out of total pairwise methylated regions)

2.3.5 Variation in DNA methylation across tissues and species

As we mentioned above, we collected both RNA and DNA from each sample in our study. We used the DNA samples to study DNA methylation patterns through low-coverage whole genome bisulfite sequencing (BS-seq). The bisulfite conversion reaction efficiency was higher than 99.4% for all samples (Table 2.2C). Following sequencing, we mapped the high-quality BS-seq reads to in silico bisulfite-converted genomes of the corresponding species and removed duplicate reads. We were able to measure methylation level in 12.5M to 22.9M CpG sites per sample, with a minimum coverage of two sequencing reads per site (Table 2.2C). We estimated local methylation levels by smoothing the data across nearby CpG sites (see Supplemental Materials; Figure 2.8; [131]). To facilitate a comparison of methylation levels across species, we annotated 10.5M orthologous CpGs in the human and chimpanzee genomes, as well as a smaller set of 2.4M orthologous CpGs in all three primate genomes (Table 2.2C-E). To identify differences in methylation levels between tissues and species we again employed a linear model framework (Methods). Focusing on methylation patterns across tissues within species, we identified between 7,026 to 41,280 differentially methylated regions between tissues, within species (T-DMRs), depending on the pairwise tissue comparisons we considered (Table 2.1B; Table 2.10A; [37]). Pairwise comparisons between hearts and lungs showed the lowest number of DMRs, regardless of species (7,026 in rhesus macaques, 8,524 in chimpanzees, 14,208 in humans), while comparisons involving heart and liver showed the largest number of DMRs (22,561 in humans, 28,767 in chimpanzee and 41,280 in rhesus macaques; Table 2.1B). We found that human T-DMRs overlapped genic and regulatory features significantly more than expected by chance. In particular, there is an enrichment of T-DMRs in intergenic regions, introns, 5'UTRs, 3'UTRs, and active enhancers (as defined by [148]; $P < 0.04$ for all tests; Table 2.10B). We found strong evidence for T-DMR conservation across all three species ($P < 10^{-16}$ across all comparisons; Table 2.11A). Though this level of conservation is higher than expected by chance, we recognize that in each tissue comparison

we performed, we had incomplete power to identify T-DMRs and so the true conservation of T-DMR is expected to be even higher. To sidestep this challenge and compare T-DMRs across species more effectively, we considered methylation data from all T-DMR orthologous regions that were classified as such in at least one species. When we performed hierarchical clustering using orthologous DNA methylation data from these T-DMRs, the data clustered first by tissue than by species (Figure 2.11). This trend is robust with respect to the species used to initially locate T-DMRs (Figure 2.11-12). Thus, our results suggest that in general, inter-tissue methylation differences within a species tend to be conserved, consistent with the observations of previous studies [92, 93, 134, 142, 153]. We next focused specifically on tissue-specific DMRs, as these may contribute to tissue-specific function. In contrast to differences in methylation between any pair of tissues, a tissue-specific DMR is defined as having a similar methylation level in three of the tissues we considered, but a significantly different methylation level in the remaining tissue. We found that there were more DMRs specific to liver (3,278 to 11,433 DMRs depending on the species) than to kidney (2,300 to 3,957 DMRs), heart (1,597 to 2,969 DMRs), or lung (453 to 5,018 DMRs, Figure 2.3; Table 2.11B). Tissue-specific DMRs are highly conserved regardless of the comparisons we made ($P < 10^{-13}$ for all comparisons, at least 25% bp overlap was required to be considered shared). In all four tissues, over 59% of conserved DMRs are hypo-methylated in a tissue-specific manner. We evaluated the overlap between tissue-specific DMRs and genomic regions marked with H3K27ac, a mark often associated with active gene expression [198]. We found that conserved hypo-methylated tissue-specific DMRs were annotated with H3K27ac more frequently than tissue-specific DMRs identified only in humans ($P < 0.001$, difference of proportions test; Table 2.11C; Supplemental Materials). We then asked about the potential impact of these conserved hypo-methylated tissue-specific DMRs on the expression of nearby genes. We found that genes with the closest TSSs to conserved tissue-specific DMRs are highly enriched with relevant functional annotations in hearts and livers (the

tissues with the largest numbers of conserved hypo-methylated tissue-specific DMRs; Figure 2.3E-F, Table 2.11D) [196]. For example, conserved heart-specific DMRs are closest to genes in cardiovascular-related pathways, including ventricular cardiac muscle cell development, canonical Wnt signaling pathway, and ERK5 cascade. Overall, these observations suggest that conserved tissue-specific DMRs are likely to underlie tissue-specific gene regulation in primates.

2.3.6 Inter-species differences in gene expression and DNA methylation levels

Our comparative catalog can be used to identify DNA methylation differences that could potentially explain gene expression differences across species and tissues. To do so, we first identified the 7,725 orthologous genes with expression data and corresponding promoter DNA methylation data in humans and chimpanzees, and the 4,155 orthologous genes with the same information for all three species. We then determined to what extent divergence in DNA methylation levels could potentially underlie interspecies differences in gene expression by comparing the gene expression effect size associated with species before and after accounting for methylation levels. To determine significant effect size differences, we applied adaptive shrinkage [171], a flexible Empirical Bayes approach for estimating false discovery rate (Methods). We note that this mediation approach does not consider the possibility that a third, unobserved event, may be causally responsible for both the methylation and expression patterns. Considering differentially expressed genes between humans and chimpanzees (in at least one tissue), we found that between 11% and 25% of genes (depending on tissue) showed a difference in the effect of species on gene expression levels once average promoter methylation levels were accounted for (significant difference in effect size classified at FSR 5% and are represented by red in Figure 2.14; Table 2.12A). As a control analysis, we considered only the genes that were not originally classified as differentially expressed

between humans and chimpanzees, and found that the difference in the effect size of species on gene expression levels was reduced in less than 1% of genes once methylation data were accounted for (FSR 5%, Figure 2.14; Table 2.12A); thus, our approach is well calibrated. We applied the same approach to the human and rhesus macaque data, and found that the percentage of genes for which gene expression differences could potentially be explained by methylation differences ranges from 21% in the lung to 40% in the liver (Figure 2.15; Table 2.12B). This observation may reflect the more extreme gene expression differences between humans and rhesus macaques than between humans and chimpanzees (prior to accounting for DNA methylation levels, P <0.003 in all tissues, t-test comparing the absolute values of the effect sizes for both groups of DE genes;). Next, we examined the genes in which DNA methylation differences may underlie inter-tissue gene expression differences (example in Figure 2.4A-C). Using adaptive shrinkage, we found that 7-25% of inter-tissue gene expression differences could potentially be explained by DNA methylation differences across tissues (FSR 5%; Table 2.12C-E). When we performed the control analysis and considered only data from genes that were not differentially expressed between tissues, less than 1% of effect sizes differed once we accounted for the methylation data (Figure 2.4F; Table 2.12C-E). Finally, we focused on regulatory patterns that are most likely to be functional; namely, conserved inter-tissue gene regulatory differences. These differences were more likely to be explained by variation in methylation levels than non-conserved inter-tissue gene expression differences (minimum difference is 7%, P <0.005 for all comparisons; at FDR <5% and FSR <5%; Figure 2.4E-4F; Table 2.12C-E). This observation is robust with respect to the FDR and FSR cutoff used (Table 2.12C-E). Indeed, the correlation between methylation and gene expression data is higher for genes with conserved inter-tissue expression patterns compared to genes whose expression patterns were not conserved (Figures 2.1E-1F). One way to maintain conserved inter-tissue expression differences could be through DNA methylation level differences. We compared the genes whose variation in inter-tissue gene expression can po-

tentially be explained by variation in DNA methylation levels (assuming no independent effect on an unobserved factor) to all genes with conserved inter-tissue expression differences (13-19% of genes across all pairwise tissue comparisons, Table 2.12C). We found that these genes are enriched for essential tissue functions (Supplemental Table 2.12F). For example, the heart genes are enriched for cardiac and smooth muscle contraction, whereas those in liver are enriched for regulation of cholesterol transport and hormone secretion (Figure 2.4G, Table 2.12F). These observations suggest that DNA methylation levels may mark or even drive differences in the expression levels in functionally relevant genes.

2.4 Discussion

2.4.1 A comparative catalog of tissue specific regulatory patterns

We designed a comparative study of gene regulation in humans, chimpanzees, and rhesus macaques that minimized confounding effects and bias. Consistent with previous studies, we found a high degree of conservation in gene expression levels when we considered the same tissue across species [16, 159, 81, 124, 176]. We also found evidence for conservation of tissue-specific DMRs. Our observations are qualitatively consistent with those of previous studies that mostly used microarrays to measure methylation levels [92, 153, 202], however, the high resolution of our sequence-based DMR data allowed us to examine a much larger number of CpG sites. Thus, we were able to show that while DNA methylation can potentially explain a modest proportion of expression differences between tissues [153], it is more likely to play a role in underlying conserved tissue-specific gene expression levels. We created and made available the most comprehensive, and likely most accurate comparative catalog of gene expression and methylation levels in humans, chimpanzees, and rhesus macaques. Comparative functional genomic studies in primates, including from our own lab, often are not designed to test for specific hypotheses. Rather, many of these comparative genome-

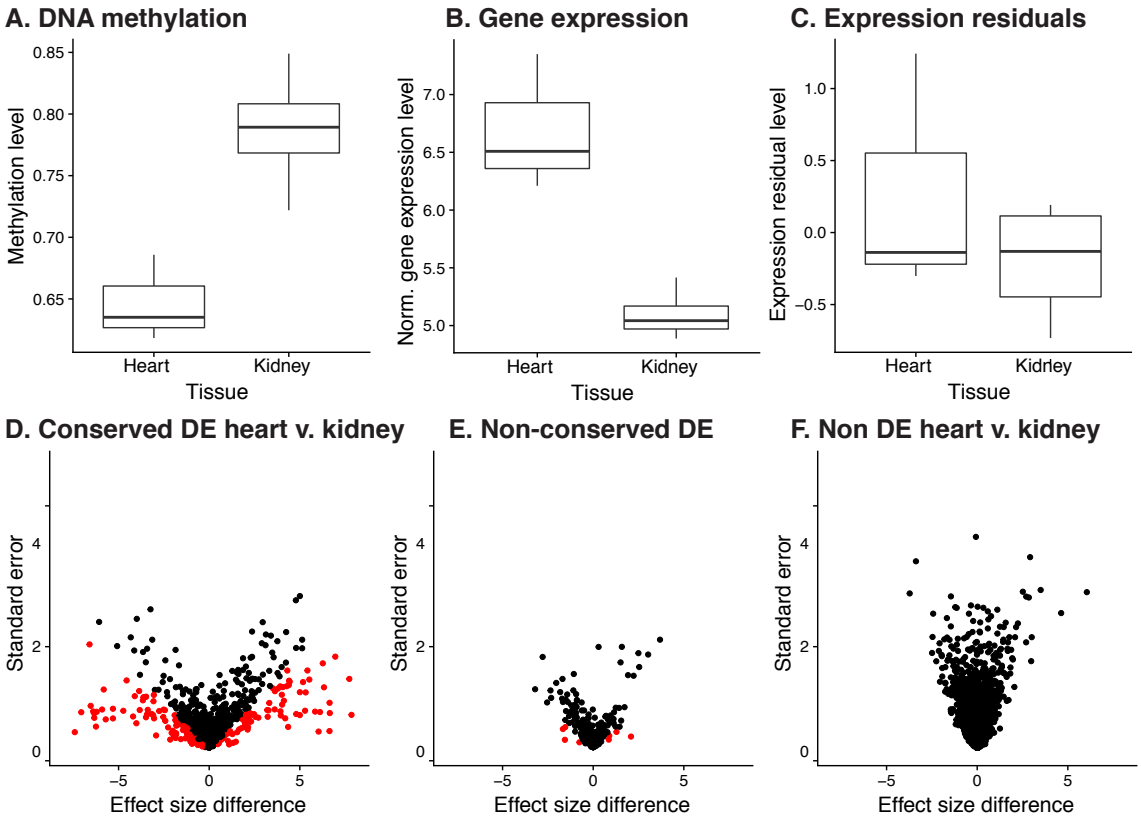


Figure 2.4: **Intertissue DNA methylation and gene expression levels (FDR = 0.05 and FSR = 0.05).** A-C. A representative example of the *PRKACA* gene in which the variation of methylation levels (A) may explain the differences in (B) gene expression levels between human heart and kidney. (C) The residuals of normalized gene expression levels after regressing out methylation levels. (D-F). Tissue effect sizes before and after controlling for DNA methylation levels in intertissue DE and non-DE genes, separately. Genes in red are significant at s-value <0.05. Effect size differences in (D) conserved DE genes in human heart relative to human kidney, (E) non-conserved DE genes in human heart relative to human kidney and (F) non-DE genes in human heart and human kidney.

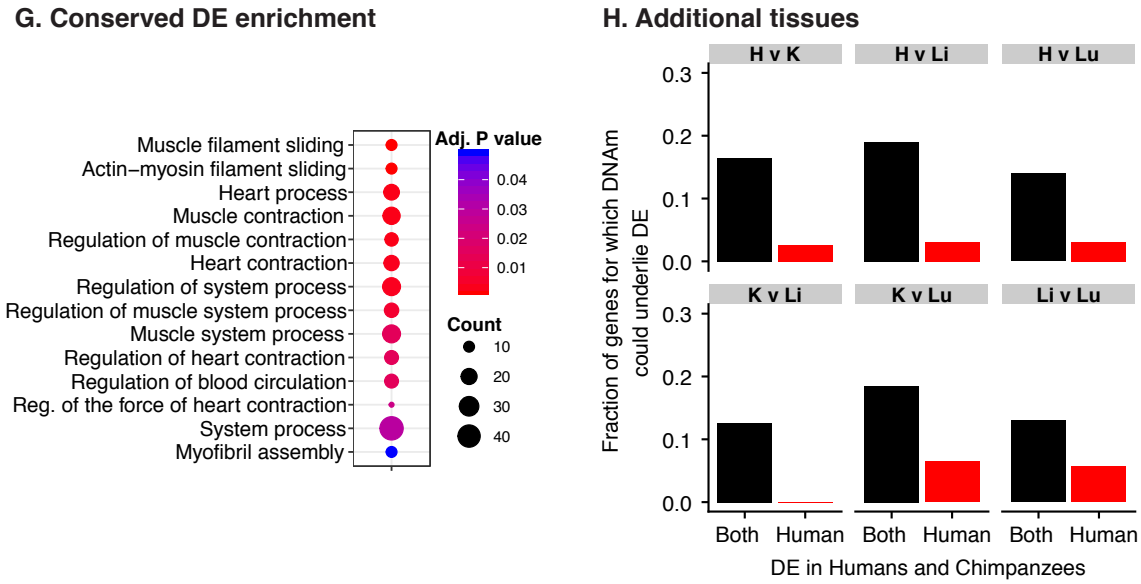


Figure 2.4, continued. (G) The conserved DE genes are enriched for heart-related function. (H) Variation in DNA methylation is more likely to explain variation in conserved DE genes than non-conserved DE genes.

scale studies aim to build catalogs of similarities and differences in gene regulation between humans and other primates. These catalogs have been shown to be quite useful; for example, they can be used to identify inter-species regulatory changes that have likely evolved under natural selection [9, 23, 24, 39, 42, 188, 84, 108, 111, 112, 115, 122, 151, 153, 177], and thereby help us better understand the evolutionary processes that led to adaptations in humans. These catalogs are also used to establish informed models of the relative importance of changes in different molecular mechanisms to regulatory evolution [113, 214], and to inform us about ancestral or derived phenotypes that may be relevant to human diseases [127, 81]. Ultimately, comparative catalogs of gene regulatory phenotypes are used to develop and test specific hypotheses regarding the connection between inter-species regulatory changes and physiological, anatomical, and cognitive phenotypic difference between species. In this study, we used a comparative catalog to identify species-specific and, in particular, tissue specific regulatory patterns, as these genes are often drug targets [59] and are likely important for the evolution of human traits [24]. We showed that genes with conserved

tissue-specific regulatory patterns have more regulatory interactions and protein-protein interactions than genes whose regulatory patterns are not conserved or are not tissue-specific. These patterns became even more pronounced when we focused on genes whose expression patterns are consistent with the action of natural selection. Put together, these observations consistently support the inference that when genes perform an important function that needs to be carefully regulated, evolution can act on multiple levels of the regulatory cascade in primates. Focusing on species-specific patterns of tissue-specific gene regulation, our observations can help formulate specific functional hypotheses regarding human-specific adaptations. For example, genes with tissue-specific gene regulation identified in humans only are enriched for GO pathways that may contribute to human-specific features, including the sodium ion import across plasma membrane in kidneys (e.g. *SLC9A3* and *TRPM4*), the glycogen biosynthetic process in livers (e.g. *PGM1* and *AKT1*), and paraxial mesoderm morphogenesis in lungs (e.g. *MST1R* and *MAP9*).

2.4.2 Consideration of study design and record keeping

Regardless of the model system used and the types of data that are collected, study design considerations are always critical. Perhaps because comparative studies in primates typically rely on opportunistic sample collection, there are no recognized study design standards that are kept and consistently reported in most existing studies (including many earlier studies from our own group). We thus believe that it is worthwhile to explicitly discuss a few important considerations regarding study design and the recording of meta-data. Without a balanced study design, it would have been impossible to independently estimate the effects of individual, tissue, and species on our data. Because the sources of confounding factors are difficult to predict in advance, we strongly recommend that samples are collected using a balanced design with respect to as many parameters as possible. These include the distribution of tissue samples per individual, the number of individuals from each species,

sex, age range, cause of death and collection time (in the case of post-mortem tissues), or sample collection and cell culturing (in the case of iPSC-based models). All steps of sample processing (RNA extraction, library, etc.) should be done in batches that are randomized or balanced with respect to species, tissue, and any other variables of interest. Most importantly, all sample processing steps should be recorded in a sample history file that includes anything that happened to any sample. We have documented many of these steps in Table 2.2A-E. This documentation can help provide evidence that a phenomenon is driven by biological rather than technical factors. It may also benefit future studies by facilitating effective meta-analysis of multiple data sets, which would help to address the problems of tissue availability and small sample sizes. We believe that, moving forward, it should be a requirement that these meta-data are available with every published comparative genomic data set.

2.5 Methods

2.5.1 Sample Description

We collected heart, kidney (cortex), liver and lung tissues from four individual donors in human (Homo sapiens, all of reported Caucasian ethnicity), chimpanzee (Pan troglodytes), and Indian rhesus macaque (Macaca mulatta), for a total of 48 samples (3 species * 4 tissues * 4 individuals; Figure 2.1A). The choice of these particular tissues was guided by their relative homogeneity with respect to cellular composition (e.g. [10]), which do not change substantially across primate species. In contrast, other tissues, such as brain subparts, differ substantially in cellular composition across primates [35], which could potentially confound the analyses. Human samples were obtained from the National Disease Research Interchange (IRB protocol 14378B). Non-human samples were obtained from several sources, including the Yerkes primate center and the Southwest Foundation for Biomedical Research, under

IACUC protocol 71619. When possible, samples were collected from adult individuals whose cause of death was unrelated to the tissues studied.

2.5.2 RNA library preparation and sequencing

In total, we prepared 48 unstranded RNA-sequencing libraries as previously described [52, 199]. Twenty-four barcoded adapters were used to multiplex different samples on two pools of libraries. RNA-sequencing libraries were sequenced on 26 lanes on 4 different flow-cells on an Illumina HiSeq 2500 sequencer in either the Gilad lab or at the University of Chicago Genomics Facility (50bp single end reads, Table 2.2; Supplemental Materials).

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

We used FastQC (version 0.10.0; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) to generate read quality report and TrimGalore (version 0.2.8), a wrapper based on cutadapt (version 1.2.1)[32], to trim adaptor sequences from RNA-seq reads. We trimmed using a stringency of 3, and to cut the low-quality ends of reads, using a quality threshold (Phred score) of 20. Reads shorter than 20 nt after trimming were eliminated before mapping (Table 2.2). For each sample, we used TopHat2 (version 2.0.8b) [116] to map the reads to the correct species' genome: human reads to the hg19 genome, chimpanzee reads to the panTro3 genome, and the rhesus macaque reads to the rheMac2 genome (Supplemental Materials). Expression level estimates may be biased across the species due to factors such as mRNA transcript size and different genome annotation qualities. To circumvent these issues, we only retained reads that mapped to a set of 30,030 Ensembl gene orthologous metaexons available for each of the 3 genomes, as described and used previously [23, 25, 82]. We defined the number of reads mapped to orthologous genes as the sum of the reads mapping to the orthologous metaexons of each gene. We quantified gene expression levels using the program coverageBed from the BEDtools suite and then performed TMM and cyclic loess normalization (Supplemental

Materials). The reason we are using hg19 is that this is still, by far, the dominant genome build in the community. In particular, all 3 releases of GTEx use hg19 and only a fraction of the 1000 genome data are available in GRCh38 coordinates. Second, to demonstrate that the results we report would not change much if we used the GRCh38 build, we leveraged the fact that differential expression analysis compares gene expression levels from groups of samples (e.g. human liver samples to human lung samples). Therefore, we compared the ranks of the normalized gene expression levels in the 15 human samples mapped using hg19 to the same samples mapped to GRCh38. The correlations of these ranks were extremely high (median Pearson's correlation = 0.96). These strong correlations suggest that our general conclusions and indeed, many genes we identified as DE would remain if we had used the GRCh38 build.

2.5.4 Analysis of Technical Variables

To assess whether the study's biological variables of interest, tissue and species– were confounded with the study's recorded sample and technical variables, we used an approach described in [21]. For the 12 RNA-seq related technical variables that were the most highly correlated with tissue or species, we assessed which technical variables constitute the "best set" of independent variables to be included in a linear model for gene expression levels. Because of the partial correlations between the variables, we applied lasso regression using the package glmnet [80]. Before performing the analysis, we also protected our variables of interest, tissue and species, in the model for each gene. We summarized each technical variable's influence across the genes by counting the number of times each technical variable was included in the "best set" of the gene models. We found that none of the technical variables appeared in more than 25% of the best sets (i.e. more than 25% of the gene models). Therefore, we chose not to include these technical variables in our model for testing differential expression. Finally, during our analysis of technical factors, we discovered that RNA extraction date was confounded with species. In 2012, we extracted RNA from the

chimpanzee samples on March 8, from the human samples on three days between March 12-29, and from the rhesus samples on March 6. To test the relationship between the date of RNA extraction and gene expression PCs in humans, we performed individual linear models on PCs 1-5 using RNA extraction date as a predictor. None of the models were statistically significant at FDR 10%, suggesting that tissue type is more highly associated with gene expression levels than RNA extraction date.

2.5.5 Differential expression analysis using a linear model-based framework

To perform differential expression analysis, we used the same approach as in [21]. We applied a linear model-based empirical Bayes method [180, 181] that accounts for the mean-variance relationship of the RNA-seq read counts, using weights specific to both genes and samples [121]. To be considered a "tissue-specific DE gene" under our stringent definition, the gene must be in the same direction and statistically significant in all pairwise comparisons including the given tissue but not significant in any comparison without that tissue. For example, for a gene to be classified as having heart-specific upregulation in a given species, the gene needed to be upregulated (a significant, positive effect size) in heart versus liver, heart versus lung, heart versus kidney, but not significantly different between the liver versus lung, liver versus kidney, or kidney versus lung, in the same species. Under the more lenient definition of tissue-specific DE genes, we compared the gene expression level of one tissue to the mean of the other three tissues. To do so, we grouped the three tissues together and again used the limma+voom framework to identify significant differences in one tissue versus the group of the other tissues. To identify inter-species differences in gene expression patterns across tissues within species (tissue-by-species interactions), we used the limma+voom framework and looked for the significance of tissue-by-group interactions. In one analysis, the groups were Great Ape versus Rhesus macaque and in another analysis, the groups were Human versus non-human primates. To minimize the number of interactions, we compared one tis-

sue relative to a group of the other 3 tissues (e.g. Great Ape versus rhesus macaque heart versus non-heart). Significant tissue-specific interactions were detected using the adaptive shrinkage method, ashr [171]. Specifically, for each test, we input the regression estimates from limma to ashr: regression coefficients, posterior standard errors, and posterior degrees of freedom. We used the default settings in ashr to calculate the shrunken regression coefficients (called the "Posterior Mean" in ashr), false discovery rate (FDR or also known as q-value), and false sign rate (FSR or also known as s-value: the probability that sign of the estimated effect size is wrong in either direction). We assigned directionality based on the sign of the posterior mean and determined significance based on the false sign rate.

2.5.6 The impact of matched tissue samples on DE results

To determine the impact of matched tissue samples on DE results, we compared intertissue DE analysis results when using tissues from the same or different individuals in GTEx v7 data [87]. We first subset the GTEx raw gene expression counts data to only individuals for which there was gene count information in the heart and lung tissues, for genes included in all 3 tissues. (There were the most GTEx samples in heart and lung; we decided to focus on these samples). Furthermore, to minimize the number of confounders needed in the linear model, we decided to only analyze individuals of the same sex and whose samples were sequenced on the same platform (sex = 1 and platform = 1 from the GTEx documentation). We then normalized the data and performed DE analysis using a voom+limma pipeline. In the linear model, we included tissue and 3 GTEx-provided covariates (covariates 1 and 2 and inferred covariate 1 from the covariate file for each tissue) as fixed effects and individual as a random effect. We chose to renormalize the raw counts data rather than use the normalized counts from GTEx because the voom+limma pipeline requires raw counts to assign voom weights. We considered the output of the intertissue DE analysis for all individuals (DE versus non-DE genes at FDR 5%) as the "ground truth". To evaluate the impact of our

study design, we then subset the gene expression information to the individuals for which there is information in all 3 tissues. We obtained gene expression level information from 4 randomly selected individuals and used the voom+limma pipeline to identify intertissue DE genes. Next, we compared the list of DE genes from this analysis to the "ground truth" list. We performed this downsampling procedure for tissues from the same 4 individuals as well as different 4 individuals 10 times each and compared the number of true and false positives from the tests. For the analysis with 8 different individuals, there were no repeated individuals, so we did not use the "duplicateCorrelation" function in voom.

2.5.7 BS-seq library preparation, sequencing, and mapping

We prepared a total of 48 whole-genome BS-seq libraries from extracted DNA as previously described [12, 200]. We aligned the trimmed reads to the human (hg19, February 2009), chimpanzee (panTro3, October 2010), or rhesus macaque (rheMac2, January 2006) genomes, and to the lambda phage genome using the Bismark aligner (version 0.8.1)[210]. We estimated the percentage of methylation at an individual cytosine site by the ratio of the number of cytosines (unconverted) found in mapped reads at this position, to the total number of reads covering this position (sequenced as cytosine or thymine, i.e., converted or unconverted) using the methylation extractor tool from Bismark (version 0.8.1). We additionally collapsed information from both DNA strands (because CpG methylation status is highly symmetrical on opposite strands [77]) to achieve better precision in methylation estimates across the genome. To obtain CpGs that mapped to multiple species, the chimpanzee and rhesus macaque CpG sites were mapped to human coordinates (hg19) using chain files from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/ and the liftOver tool from the same website. These files had previously been filtered for paralogous regions and repeats, but we also removed positions that were not remapped to their original position when we mapped from human back to their original genome. Chimpanzee and rhesus macaque CpG

sites were mapped to human, even if their orthologous positions were not a CpG site in human.

2.5.8 Identifying differentially methylated regions (DMRs)

We were interested in identifying regions exhibiting consistent differences between pairs of tissues or pairs of species, taking biological variation into account. To identify DMRs we used the linear model-based framework in the Bioconductor package bsseq (version 0.10.0)[88]. For a given pairwise comparison (e.g., human liver vs. human heart), the bsseq package produces a signal-to-noise statistic for each CpG site similar to a t-test statistic, assuming that methylation levels in each condition have equal variance. As recommended by the authors of the package, we used a low-frequency mean correction to improve the marginal distribution of the t-statistics. Similar to previous studies using this methodology, a t-statistics cutoff of -4.6,4.6 was used for significance, [88]. DMRs were defined as regions containing at least three consecutive significant CpGs, an average methylation difference of 10% between conditions, and at least one CpG every 300 bp [88]. We used BEDTools (version 2.26.0) [160] to calculate the number of overlapping DMRs across tissues and/or species [37]. We required overlapping DMRs to have a minimum base pair overlap of at least 25%, unless otherwise stated. To be considered a tissue-specific DMR, the region was required to be a significant tDMR in 1 tissue compared to the other 3 tissues pairwise (in the same direction) but not a significant tDMR across any of the other 3 tissues in pairwise comparisons. We again used BEDTools to ensure a minimum base pair overlap of at least 25%. Once the tissue-specific DMRs were identified within a species, we then classified them as species-specific or conserved. To be considered conserved (across humans and chimpanzees or all three species), the tissue-specific DMR had to be significant in all species in the comparison and have a minimum base pair overlap of the tDMR of at least 25%.

2.5.9 Calculating the average methylation levels of conserved promoters

To calculate the DNA methylation levels of orthologous CpGs around the transcription start site (TSS) of orthologous genes, we first had to determine the orthologous TSSs. We began with the 12,184 orthologous genes in our RNA-seq analysis. Of these, we found that 11,131 of these orthologous genes had an hg19 RefSeq TSS annotation. We used liftOver to find orthologous sites in the chimpanzees and rhesus macaque genomes in 9,682 of those 11,131 genes. We then determined which of the hg19 RefSeq TSS annotations were closest to the first hg19 orthologous exon, and repeated this process with the other two species and their respective genomes. We found that 9,604/9,682 of the closest TSS annotations in humans at the same liftOver coordinates in the other two species. We then calculated the distance between the first orthologous exon to the TSS site in all 3 species individually. To minimize this difference between the 3 species, we filtered all genes with a maximum distance difference across the species of larger than 2,500 bp. (For reference, the 75th percentile of the maximum difference in distance was 2,078 bp.) 7,263 autosomal genes remained after this filtering step. 4,155 genes had at least 2 orthologous CpGs 250 bp upstream and 250 bp downstream of the orthologous TSS. We chose a 250 bp window around the TSS based on DNA methylation levels around the promoter in [12] and calculated the average of orthologous CpGs within this window for the 4,155 genes. Using the same method but in humans and chimpanzees only, we found and calculated the average of orthologous CpGs within this window for 7,725 genes.

2.5.10 Joint analysis of promoter DNA methylation and gene expression levels

To determine whether DNA methylation may underlie interspecies differences in gene expression levels, we used a joint analysis method as described below. For each gene, we analyzed the gene expression levels, along with the accompanying average methylation level

250 bp upstream and downstream of the TSS (found above). For a given tissue, we first determined the effect of species on gene expression levels using a linear model, with species and RIN score as fixed effects (Model 1). Next, we parameterized a linear model attempting to predict expression levels exclusively from methylation levels. We refer to these residuals as "methylation-corrected" gene expression values. We then used these values to again determine the effect of species, this time on gene expression levels corrected for methylation, using a linear model with species and RIN score as fixed effects (Model 2). To determine the contribution of DNA methylation levels to inter-species differences in gene expression, we computed the difference in the species effect size between Model 1 and Model 2 for each gene, as well as the standard error of the difference. Large effect size differences between Models 1 and 2 for a given gene suggest that methylation status may be a significant driver of DE. To assess the significance of this difference, we used adaptive shrinkage (ashr) [171] to compute the posterior mean of the differences in the effect sizes, using vashr, with the degrees of freedom equal to the number of samples in the linear model minus 2. The shrunken variances from vashr were used in the ashr posterior mean computation. From this procedure, we obtained the number of genes where species has a significant difference in effect sizes before and after regressing out methylation. We assessed significance using the s-value statistic (false sign rate, FSR [171]). Using the s-values, rather than the q-values, not only takes significance into account but also has the added benefit of assessing our confidence in the direction of the effect. We performed the above analysis separately for inter-species DE genes and non-DE genes, and in each tissue individually. We identified inter-species DE genes in our tissue of interest as those with a significant species term in the model of species and RIN score as fixed effects. We assessed significance of DE genes at FDR 5%, unless otherwise noted. We also applied the same analysis framework to determine whether DNA methylation may underlie inter-tissue differences in gene expression levels. For the inter-tissue DE genes and non-DE genes, we replaced "species" with "tissue" as a fixed effect in

models 1 and 2. We assessed significance with various FDR and FSR thresholds, as specified in the text.

2.5.11 Data and code availability

All raw and processed sequencing data generated in this study have been submitted to the NCBI's Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo) using GEO Series accession number GSE112356. Data and scripts used in this paper are available at https://github.com/Lauren-Blake/Reg_Evo_Primates. The results of our scripts can be viewed at https://lauren-blake.github.io/Regulatory_Evol/analysis/.

2.6 Acknowledgments

Members of the Gilad, Marques-Bonet, Robinson-Rechavi, Stephens, and Pritchard labs provided helpful discussions and comments on the manuscript. In particular, Matthew Stephens provided guidance on integrating the DNA methylation and gene expression data, and Michelle Ward and Natalia Gonzales provided helpful comments on a draft of the manuscript. We thank Kasper Hansen, Jenny Tung, and Luis Barreiro for discussions regarding multispecies methylation analysis. We acknowledge Athma Pai for sharing a methylation protocol, Bryce van de Geijn for help with the WASP pipeline, Charlotte Soneson, Jacob Degner, and Unjin Lee. We also thank the Yerkes Primate Center and Southwest Foundation for Biomedical Research, Anne Stone and Jssica Hernndez Rodrguez for providing and/or helping to ship the tissue samples. L.E.B. was supported by the National Science Foundation Graduate Research Fellowship (DGE-1144082). Additionally, L.E.B. and I.E. were funded by the Genetics and Gene Regulation Training Grant (T32 GM07197). J.R. was funded by a Swiss NSF postdoc mobility fellowship (PBLAP3-134342) and the Marie Curie International Outgoing Fellowship PRIMATE_REG_EVOL. This project was funded in part by the ORIP/OD P51OD011132 grant. T.M.B. is supported by BFU2017-86471-P

(MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya. The content presented in this article is solely the responsibility of the authors and does not necessarily reflect the official views of the funders. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.6.1 Author contributions

TMB and YG conceived of the study and designed the experiments. JR and NEB performed the experiments. CC extracted the RNA and prepared the sequencing libraries. LEB, JR, IHH, and RGP analyzed the results (with input from CJH and IE). YG supervised the project. LEB and YG wrote the paper. All authors read and approved the final manuscript.

2.7 Supplementary Information

2.7.1 Supplemental Methods

RNA library preparation and sequencing A relatively small percentage of reads could not be assigned to any sample because their adaptor sequence did not match any of the adaptors used in the study. Therefore, we ran in-house Perl scripts to recover those reads that differed at a single position. Because of a calibration issue with the Gilad lab Illumina HiSeq sequencer, the first and the third flow-cell of the study (16 lanes) yielded a low number of reads. However, this problem did not affect the quality of the reads, so we kept these lanes for the analysis.

Quantifying the number of RNA-seq reads from orthologous genes When mapping to the species' genome, we allowed for up to two mismatches in each read and kept only reads that mapped uniquely. Tophat2 uses unmapped reads to perform gapped alignments to the

genome and discover new exon-exon junction sites. For this step, we disabled the coverage-based search, and only 1 mismatch was allowed in the anchor region of the reads (>=8 nt). The minimum intron length was set to 70 nt and the maximum to 50000 nt. This yielded from 28,544,039 (R1H) to 72,808,273 (R4Li) mapped reads across samples (mean = 46,514,692 reads). Mapping rates were between 71% and 94%. We performed all downstream analyses in R (versions 3.1.1, 3.2.2, or 3.4.3) unless otherwise stated.

RNA-seq data transformation and normalization We calculated the log2-transformed counts per million (CPM) from the raw gene counts of each sample using edgeR [165]. We then filtered out lowly expressed genes, keeping only genes with an expression level of log2(CPM) >1.5 in at least 24 of the 48 samples [166]. We normalized the original read counts using the weighted trimmed mean of M-values algorithm (TMM) [166]. This process helped us to account for differences in the read counts at the extremes of the distributions. We then calculated the TMM-normalized log2-transformed (CPM) values for each of the genes. After performing normalization, we performed principal components analysis (PCA) using the TMM-normalized log2-transformed (CPM) values of all genes, 1 human heart sample (H1H) clustered with the human livers rather than the hearts. After performing SNP calling on the RNA-seq data (see section below), we found that the SNPs in sample H1H matched those from the other tissues from this individual. We removed this sample (H1H) from the list of the original gene counts. We again filtered for lowly expressed genes keeping only genes with an expression level of log2(CPM) >1.5 in at least 24 of the 47 samples. We also wanted to allow for small differences in the distributions of gene expression across tissues. Therefore, on the 12,184 remaining genes, we performed a TMM normalization and then performed a cyclic loess normalization with the function normalizeCyclicLoess from the R/Bioconductor package limma [11, 164]. To run PCA, we used the R function prcomp. For hierarchical clustering, we used unsupervised agglomerative clustering on the correlation matrix of the gene expression data. In this transformation and normalization process, we were interested in the impact

of sample-specific biases in GC content on the gene expression counts. Therefore, we used the WASP pipeline [205] to obtain expected GC-normalized counts. Specifically, we filtered the genes with lowly expressed counts so that only genes with the log2-transformed counts per million (CPM) >-5.5 in at least 2 of the 4 samples in each species-tissue pair (e.g. 2/4 chimpanzee hearts) remained. For each of the 16,616 genes that remained, we summed the read depth (raw counts) of the 4 samples in each tissue-species pair. We used the WASP pipeline [205] (https://github.com/bmvdgeijn/WASP/blob/master/CHT/update_total_depth.py) to obtain expected read counts, adjusted for read depth and GC content. The GC content for each of the orthologous metaexons was previously calculated as part of [82]. For each tissue-species pair, the adjusted raw counts and the actual raw counts were highly correlated (¿ 0.98). Therefore, we did not adjust for read depth or GC content in our RNA-sequencing data.

SNP calling in the RNA-seq and BS-seq data We called single nucleotide variants on RNA-seq data from each tissue and sample using standard hard filtering parameters according to GATK recommendations [206]. Briefly, duplicated reads were removed using Picard MarkDuplicates (http://broadinstitute.github.io/picard). Reads were then subjected to local realignment, base-score recalibration, and candidate-variant calling using the IndelRealigner, TableRecalibration, and HaplotypeCaller tools from GATK [137]. We required a base quality score >20. We only considered variants that were observed in at least four of the samples. We used the same method for the BS-seq data. Through this process, we found that 2 groups of samples had been mislabeled during sequencing: the sample labelled R3Li was actually R2Li, and R3Lu was actually R2Lu.

Analysis of Technical Variables We recorded variables related to the samples (e.g. sex), variables specific to gene expression (e.g. RNA-seq flow cell number), and variables related to methylation levels (e.g. number of CpG sites covered) (Table 2.2). Briefly, we determined which of our recorded technical variables were significant predictors for each of the gene

expression PCs 1-5 using individual linear models for each of the gene expression variables (FDR ¡ 10% for each test). The significant technical variables were then tested against our biological variables of interest, tissue and species, again with individual linear models. For the numerical technical variables, we quantified the strength of these associations using the P values from analysis of variance (ANOVA), and used a Chi-squared test (using Monte Carlo simulated P values) for the categorical technical variables (significance at FDR <10%). We repeated the same analysis for methylation data, testing the associations between methylation PCs 1-5 and sample information.

Differential expression analysis using a linear model-based framework We implemented linear model-based framework using the R packages limma and voom [121, 180, 181]. This pipeline has previously been shown to perform well with at least 3 samples per condition [162, 185]. We hypothesized that RNA quality may be impacted by postmortem time prior to collection. According to the documentation that we received from the different sites, all of the rhesus macaque samples were collected earlier than all of the chimpanzee samples. These differences could impact RNA quality. Hence, we used RIN score as a proxy for RNA quality, and included RIN score in the linear models. In the linear models, species, tissue, RIN score, and species-by-tissue interaction terms were modeled as fixed effects. Individual was modeled as a random effect. We used contrast tests in limma to identify genes that were differentially expressed between tissues within each species and across species in the same tissue. We corrected for multiple testing with the Benjamini and Hochberg false discovery rate (FDR) [18]. Genes were considered significantly DE at FDR-adjusted P values <0.01, unless otherwise stated.

Comparing the rank of tissue-specific DE genes in our dataset to the Genotype-Tissue Expression (GTEx) Project To benchmark the conserved tissue-specific DE genes, we compared the rank of the gene expression level in our data to the rank of the gene expression level in the GTEx v6 heart, liver, lung, and kidney tissue data [87]. After this comparison, we looked

for enrichment of genes with a given rank. To do so, we used the R package topGO [1], with the same implementation as in [26]. This implementation included the use of Fisher's Exact Test, with topGO's weight01 algorithm (which takes into account the correlation among GO categories within the graph structure of the program). We then repeated this process for the tissue-specific DE genes identified in humans only.

The expected overlap and significance of the observed overlap We used the process from [153], based on the hypergeometric distribution, to assess the expected overlap of the conserved DE genes and significance of the observed number of conserved DE genes. This process relies on comparing a population proportion to a sample proportion. We first asked about the overlap of DE genes in humans and chimpanzees. To be conservative, in the case of the human and chimpanzee overlap, we assigned the species with the greater number of genes in the direction of interest as the population and the other species as the sample. To assess the expected overlap in upregulated human and chimpanzee genes in a given tissue, we used the P value from the hypergeometric distribution with the following parameters: m is the total number of DE genes in the population, n is the total number of upregulated genes in a population minus m, q is the observed overlap of upregulated DE genes (between the humans and chimpanzees), and k is the total number of upregulated DE genes in the sample, all within the given tissue. The "expected overlap" is the value at which the maximum likelihood estimate for which m, n, and k occurs. We then repeated this process for the upregulated and downregulated DE genes, in all 4 tissues separately. To obtain the same statistics for the tissue-specific DE genes, we used only the tissue-specific DE genes, and calculated n as the total number of genes upregulated in the tissue of interest compared to the other three tissues in the population minus the number of tissue-specific DE genes in the population. To calculate these statistics for all 3 species, we used this framework to ask whether the observed overlap between all three species was significant relative to the overlap of the human and chimpanzee DE genes. In the same manner, we used the hypergeomet-

ric distribution to assess the expected overlap and significance of the number of conserved tDMRs (tissue differentially methylated regions) and conserved tissue-specific DMRs.

The overlap between DE genes and previously defined networks

To find gene expression patterns that are consistent with the action of natural selection, for each significant DE gene, we determined the within-species variance of all 3 species and found the average of the 3 variances. We then ranked the genes by the mean variances. For the co-transcription network analysis, we used the shared TE-TE networks for the heart and lung as well as the heart specific network from the supplementary materials in [172]. We downloaded the list of protein-protein interactions in the heart, kidney, liver, and lung from the Human Protein Atlas [204]. For the interactions analyses, we counted the number of interactions for each gene in the co-transcription networks or the protein-protein interactions list, in the appropriate tissue.

BS-seq library preparation, sequencing, and mapping To assess the efficiency of the conversion reaction [27], we spiked the extracted DNA with unmethylated lambda phage DNA. For each sample, we prepared at least two libraries, with independent PCR amplifications to minimize PCR duplication rates. The BS-seq libraries were sequenced on 111 lanes on 17 flow-cells on an Illumina HiSeq 2500 sequencer in the Gilad lab or at the University of Chicago Genomics Facility (Table 2.2). Reads were single-end and 49 to 59 bp. The distribution of libraries of technical replicates over the flow-cells is described in Table 2.2. Similar to RNA-seq data, we used FastQC to generate quality reports. TrimGalore (version 0.2.8) was used to trim adapter sequences incorporated in the BS-seq reads, using a stringency of 3, and to cut the low-quality ends of reads, using a quality threshold of 20. We eliminated reads shorter than 15 bp post-trimming. We aligned the trimmed reads to the human (hg19, February 2009), chimpanzee (panTro3, October 2010), or rhesus macaque (rheMac2, January 2006) genomes, and to the lambda phage genome using the Bismark aligner (version 0.8.1)[210]. The Bismark aligner maps reads to in-silico converted (G to A and C to T)

genome sequences using Bowtie (version 1.0.0). This aligner was shown to perform well on benchmark studies [18, 163, 203]. We permitted one mismatch in the seed of the alignment, and by default Bismark reports only uniquely mapped reads. Across technical replicates, mapping rates ranged from 49% to 82% (median 76%; Table 2.2). We applied the Bismark deduplication script to each technical replicate to remove reads mapped to the same starting genomic position, which likely arise through PCR amplification of the same DNA fragments during library preparation [27]. Across technical replicates, the duplication rates ranged from 2.8% to 44% (median 11%; Table 2.2). To determine the bisulfite conversion efficiency, we calculated the conversion rate at cytosines from the spiked-in lambda phage DNA (for which coverage ranged from 12x to 107x). We found this rate to be at least 99.4% across technical replicates (Table 2.2), increasing confidence in our data. After combining all technical replicates, the average genome-wide coverage, calculated at CpG sites shared across the three species, ranged from 1.7 to 5.7 per sample (median of 4), corresponding to 12M to 23M CpG sites with a coverage of at least 2 (median 19M; Table 2.2).

Methylation level estimate smoothing Since methylation data from BS-sequencing is typically lower coverage than methylation array data, we first applied a smoothing procedure on raw methylation levels for each sample.It has previously been shown that this procedure increases the precision of low-coverage BS-seq data, and yields methylation estimates that are in excellent agreement with high-coverage BS-seq data without smoothing [88]. To perform the smoothing, we used the BSmooth method (as implemented in the Bioconductor package bsseq, version 0.10.0) [88]– with the default parameters for smoothing– at least 70 CpG sites with methylation data in a smoothing window of at least 1 kb. We had 17.6M human-chimpanzee orthologous CpGs and 7.5M CpGs orthologous across all 3 species. Next, we filtered the orthologous CpGs based on coverage to increase confidence in our data. We eliminated sites with >10x coverage, as the relative sparsity of the data suggests that these reads were likely mapped to repeated regions. For each CpG site in each species/tissue

combination, we required an average of 2x coverage and that at least 2 out of 4 individuals had a coverage >= 2x. After filtering based on coverage, we had 2.4M autosomal CpGs orthologous in all three species (10.5 million in humans and chimpanzees only).

Identifying differentially methylated regions (DMRs) For a given pairwise comparison (e.g., human liver vs. human heart), the bsseq package produces a signal-to-noise statistic for each CpG site similar to a t-test statistic, assuming that methylation levels in each condition have equal variance. As recommended by the authors of the package, we used a low-frequency mean correction to improve the marginal distribution of the t-statistics. Similar to previous studies using this methodology, a t-statistics cutoff of -4.6,4.6 was used for significance [88].

Overlap of tissue-specific DMRs with regulatory regions We extracted the coordinates of the following features from GENCODE annotation (release 19)[89]: exons, first exons, CDS, 3' and 5' UTRs, introns, intergenic regions, promoters (2 kb around the TSS [66, 226]) and proximal promoters (250 nt around the TSS [49]) of all genes, or only of protein coding genes. We downloaded coordinates of the CpG islands from the UCSC Genome Browser [109]. CpG islands (identified as segments of the genome with %G andC >50%, length >200 nt, and a ratio of observed over expected number of CG dinucleotides based on number of Gs and Cs in the segment >0.6). CpG island shores were defined as 2 kb regions flanking CpG islands, and CpG islands shelves were defined as 2 kb regions outside of CpG islands shores [56]. To test the overlap of tissue-DMRs with enhancers, we used the set of tissue-specific enhancers defined by the FANTOM consortium using CAGE-seq data on primary tissues [148]. To control for the potential effects of CpG density and region length in these analyses, we generated 100 sets of randomly located control regions matching the length and CpG densities of the DMRs in the studied set. We calculated the overlap of tissue-specific DMRs with H3K27ac in the left ventricle of the heart, kidney, liver, and lung adult tissues from the Epigenome Roadmap (The Roadmap Epigenomics Consortium 2015). We

used the consolidated broad peak data for the left ventricle, liver, and lung (available from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/broadPeak/). Since a consolidated version was not available for the kidneys, we used unconsolidated kidney sample numbered 153 (available from http://egg2.wustl.edu/roadmap/data). We used BEDTools (version 2.26.0) [160] to calculate the number of tDMRs that overlap an H3K27ac mark.

2.7.2 Supplementary Text

Assessing the impact of technical variables on gene expression levels and methylation levels
We tested the relationship between our technical factors and biological variables of interest, namely tissue and species (see Methods). Through this process, we discovered that RNA extraction date was confounded with species (Figure 2.5B). In subsequent analysis with only the human samples (as humans were the only species with samples processed on multiple days), we found that the RNA extraction date did not highly correlate with tissue (Methods, Figure 2.5C). Given our particular interest in tissue differences, we do not think that differences in RNA extraction date had a larger impact on variation in gene expression levels than tissue type. Furthermore, the time of tissue collection post mortem is also confounded with species (Table 2.2A). These differences could impact RNA quality, which can be approximated by RIN score. Indeed, RIN scores were typically higher in rhesus macaques than the other species (Table 2.2B; Figure 2.5A). As a result, we included RIN score as a covariate when modeling gene expression levels. The RNA quality may have impacted the number of differentially expressed (DE) genes identified between tissues. This pairwise DE was higher in rhesus macaque than in chimpanzee and human across FDR cutoffs (Table 2.6B). This finding is potentially due to the higher sample quality, and therefore lower gene expression level variance, in rhesus macaques. We also tested for associations between technical factors and biological variables of interest in the BS-seq data. Most of the significant associations

were related to DNA methylation levels (e.g. number of orthologous CpGs sites with low methylation, mean methylation level at orthologous CpGs) across species and tissues. We expect the DNA methylation level densities to vary somewhat across tissues [153] and therefore, these inter-tissue differences are likely biological rather than technical. Additionally, we found a slightly higher overall methylation level in human, compared to chimpanzee and macaque samples (average in humans = 0.664, in chimpanzees = 0.646, in rhesus macaques = 0.626; Figure 2.8), which persisted in the raw, unsmoothed data (Figure 2.8A). The distribution of methylation levels could potentially be biased by CpG to TpG homozygous or heterozygous SNPs, which are erroneously inferred as unmethylated CpG sites. If the rate of such SNPs was higher in the studied chimpanzee or macaque individuals compared to human (with respect to their respective reference genome), we could observe differences between species. Since we had performed SNP calling on our RNA-seq dataset, we retained only CpG sites located in orthologous exons, and excluded sites with C to T SNPs in any of the samples. However, we still observed differences between species (Figure 2.8C). Moreover, higher methylation rates in humans compared to chimpanzees were previously reported (but rarely discussed) in a diverse set of tissues, using various technologies to measure methylation [83, 93, 161, 142]. Therefore, this result may be driven by biases when mapping to different species' genomes. The coverage on the lambda phage genome was statistically significant (FDR $<10^{-10}$). However, further analysis showed that this trend was driven by some low values in the chimpanzee samples, rather than the rhesus macaques (Table 2.2D). Indeed, there is no difference in this factor across the human and the rhesus macaque samples (P = 0.15, Student's t-test). Therefore, we do not think that coverage on the lambda phage genome can account for the differences in DNA methylation between the Great Apes and the rhesus macaque samples (Figure 2.1D). We found evidence for a dependent relationship between species and lane number (Chi squared test, FDR= 10^{-13}). Since these lanes were spread across multiple flowcells, we do not think that lane substantially contributed to the

variance in DNA methylation levels. We also found evidence for a dependent relationship between tissue and library preparation date (Chi squared test, FDR = 0.006). Since the correlation was modest (Pearson's correlation = -0.22) and most tissues within a species had libraries made on multiple days, we chose not to correct for this variable. Finally, we note that sample age has previously been shown to impact methylation status in a subset of genes [57, 100]. DNA methylation levels were weakly positively correlated with age (age quantile relative to the species' average lifespan; Pearson correlation's = 0.18). However, in our factor analysis, this relationship was non-significant (FDR >10%). Therefore, we did not correct for this variable.

Tissue-specific expression patterns We asked whether the tissue-specific expression patterns we found in a sample of four individuals from each species are indeed indicative of regulatory patterns in a larger population. To examine this, we again considered human GTEx data from the same four tissues we included in our study (see Methods). Because the sample size of the GTEx data is much larger than in our study, the quantity we compared between our data and that from GTEx is the normalized gene expression ranks in the four tissues. For example, in both our and the GTEx data, troponin T2 (*TNNT2*) shows the highest expression in the heart (rank 1) and the lowest expression in the liver (rank 4). Using this approach, we found that 428 (62%) of the 687 genes with a tissue-specific expression pattern exclusively in our human data have the same tissue-based ranked expression in the GTEx data. This observation suggests that tissue-specific expression patterns found in just four individuals are quite often not representative of the regulatory patterns in the larger population. In contrast, however, we found that 1,530 (88%) of the 1,739 genes with a conserved tissue-specific expression pattern based on our data have the same tissue-based ranked expression in the GTEx data. Thus, conserved differential expression significantly increases the confidence of classifying tissue-specific expression patterns in a larger human population ($P < 10^{16}$, difference of proportions test).

Adaptive shrinkage and false sign rate to identify tissue-specific genes We investigated to what extent our ability to detect tissue-specific genes could be substantially impacted by differences in effect sizes across the species (Table 2.7). Therefore, we tested the use of an adaptive shrinkage method [171] to identify genes with a small effect size but consistent direction of effect in each species and used the accompanying false sign rate (FSR) instead of FDR thresholds. The percentage overlap was relatively robust to threshold method (Table 2.7). We found that this method increases both the total number of tissue-specific differences and the species-specific differences. However, it also increases the number of conserved tissue-specific gene expression differences in humans and chimpanzees relative to those in chimpanzees and rhesus macaques (Table 2.7), more closely reflecting established phylogenetic relationships.

Identifying inter-species differences between tissues Since our data contained multiple tissues and species, we identified genes with interspecies differences between tissues (tissue-by-species interactions). These tissue-by-species interactions are potentially informative for Great Ape evolution (when the contribution of species on gene expression in a given tissue is different between Great Apes and rhesus) and the evolution of human-specific mechanisms in tissues (when the effect of species on gene expression in a given tissue is different between humans and a group containing chimpanzees and rhesus macaques). Using a linear-model based framework, we modeled these differences with tissue-by-species interaction terms (Methods). We found 664 total significant interactions in the Great Ape versus rhesus macaque comparison and 91 in the human versus chimpanzee and rhesus macaque (FDR 1%; Table 2.9). Given our sample size and the small effect sizes of these interactions, we are probably underpowered to detect such interactions. To address this, we employed an adaptive shrinkage method [171] to identify genes with a small effect size but consistent direction of effect in each species and used the accompanying false sign rate (FSR) instead of FDR thresholds. This method was used to identify cases where the observed sign of the effect across tissues

was different between species. After applying this method, we found 1,006 Great Ape-by-tissue interactions and 257 human-by-tissue interactions (FSR = 1%; Table 2.9). Potentially the most interesting class of tissue-by-species interaction is when species impacts one tissue differently than the other three tissues. Therefore, we used ASH to find 799 Great Ape-by-tissue interactions and 249 human-by-tissue interactions only present in one tissue (FSR = 1%; Table 2.9). We defined tissue-by-species specific interactions as interactions with an effect size sign different from the signs of the other interactions (e.g. a positive sign when all other signs are 0 or negative). Unsurprisingly, even after accounting for small effect sizes, there were more tissue-by-species interactions for Great Apes versus rhesus macaques than human-specific ones.

Promoter DNA methylation quality To check our promoter DNA methylation levels in the humans and chimpanzees, we subset the DNA methylation promoter data to the 3 human and chimpanzee tissues tested in a previous study from our lab [153]. Consistent with this previous study, PC1 was more highly correlated with tissue than species and PC2 was more highly correlated with species than tissue (Figure 2.10A). Even in this subset of the data, there was more clear separation between tissues in the gene expression levels than the promoter DNA methylation data for these genes (Figure 2.10B).

Identification of differentially methylated regions across species (S-DMRs)

Using the same method to identify DMRs across tissues, we then identified thousands of DMRs across species (sDMRs). We found the lowest number of sDMRs on autosomal chromosomes in lungs (8617 DMRs between human and chimpanzees, 17696 DMRs between humans and rhesus macaques, and 15544 between chimpanzees and rhesus macaques) and highest total number in hearts (14504 DMRs between human and chimpanzees, 25539 DMRs between humans and rhesus macaques, and 15544 between chimpanzees and rhesus macaques, Table 2.1B). Similar to the pairwise DE analysis across species, the number of DMRs between species are consistent with known phylogenetic relationships. However, un-

like in the pairwise DE analysis across species, the number of S-DMRs is sometimes higher than the number of pairwise T-DMRs. For example, there are more lung S-DMRs than human heart-lung DMRs. This trend is somewhat unexpected given the gene expression data, but consistent with clustering pattern of the methylation data (Figure 2.1D).

2.8 Supplementary Figures

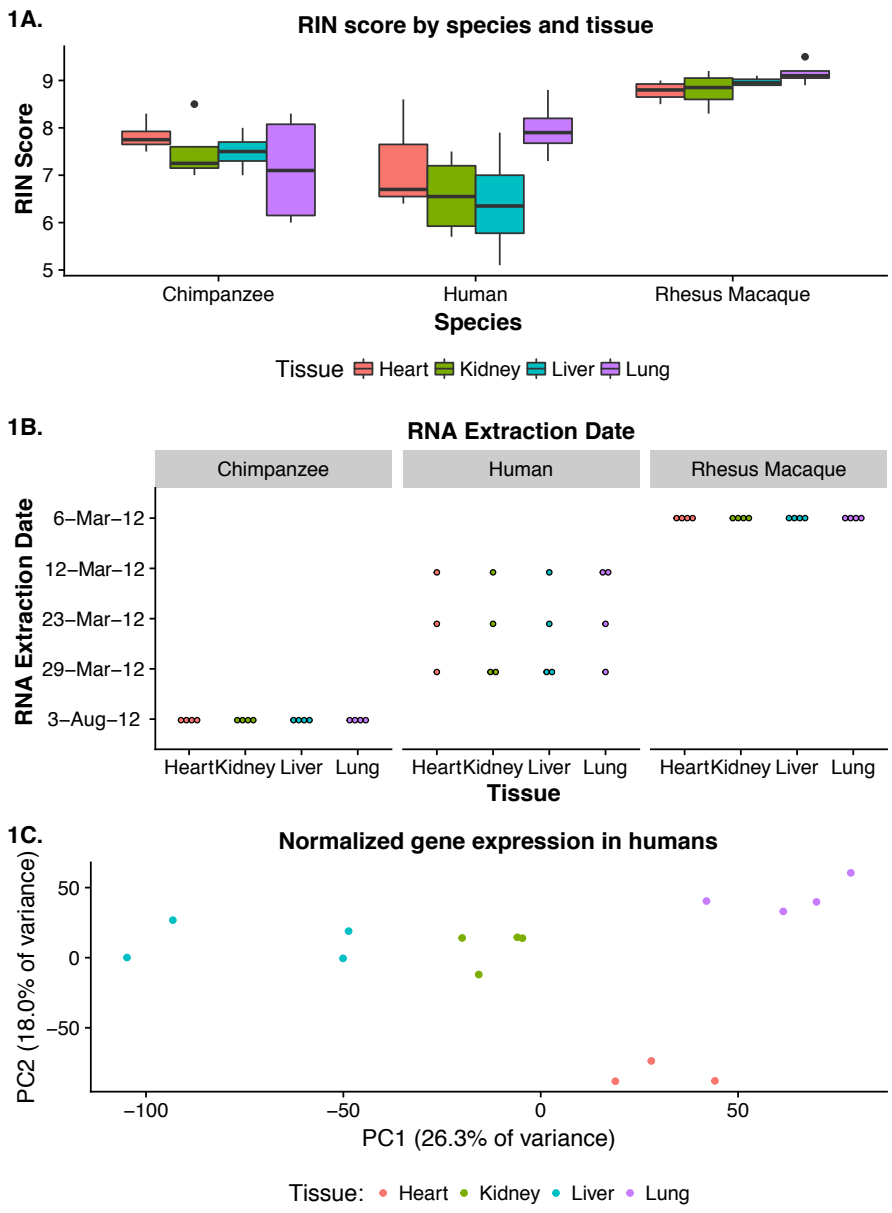


Figure 2.5: **Distributions of potential confounders across biological variables of interest.** (A) RIN score across the samples. (B) RNA extraction date by species (C) PCA of RNA extraction date in humans.

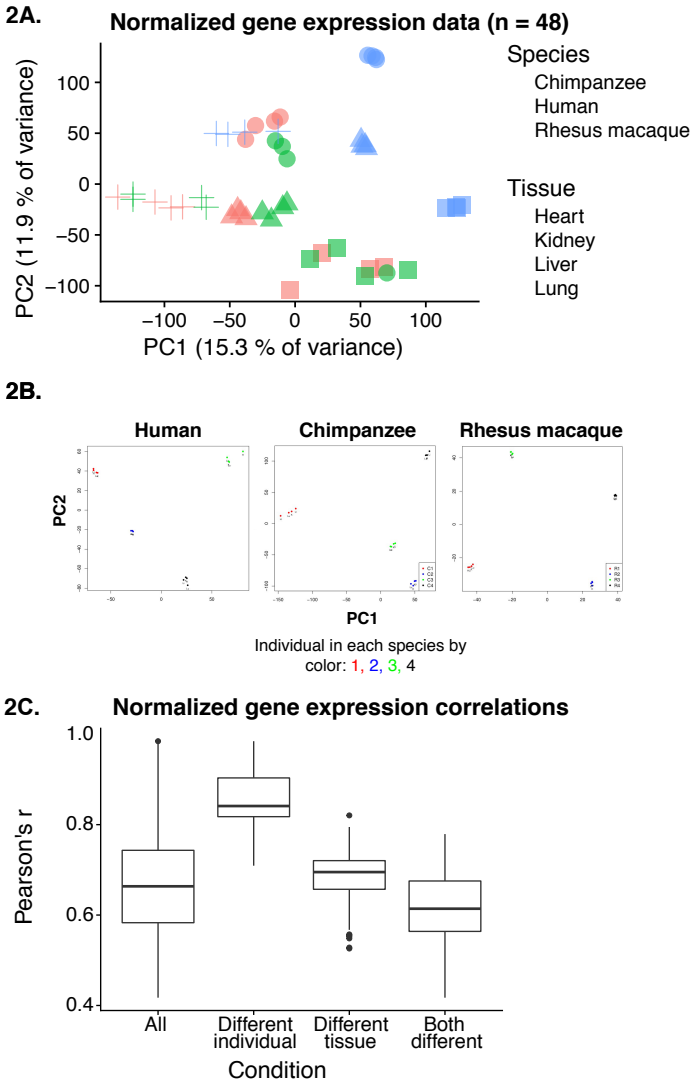


Figure 2.6: **Sample QC.** (A) One human heart (orange triangle) clusters with the human livers (teal triangles). The sample originally labeled as "human 1 heart" is likely a liver from the same human. (B) GATK analysis of the sample labelled "human 1 heart" clusters with the human 1 liver. (C) Gene expression levels from different tissues in the same individual ("different tissue") are more highly correlated than gene expression levels between tissues from different individuals ("both different") and all combinations of tissues and individuals ("all"). Within a given tissue, gene expression levels are most highly correlated ("different individual").

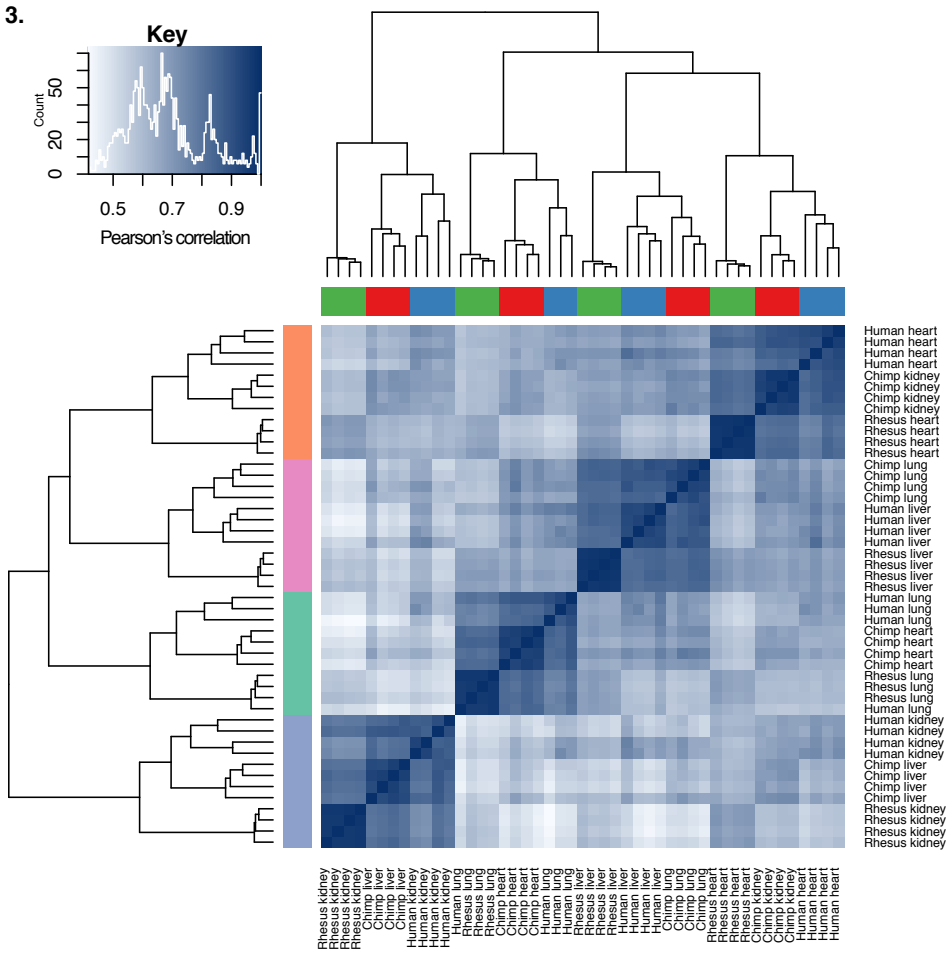


Figure 2.7: **Correlation matrix of normalized log2(CPM) gene expression values from 12,184 genes.**

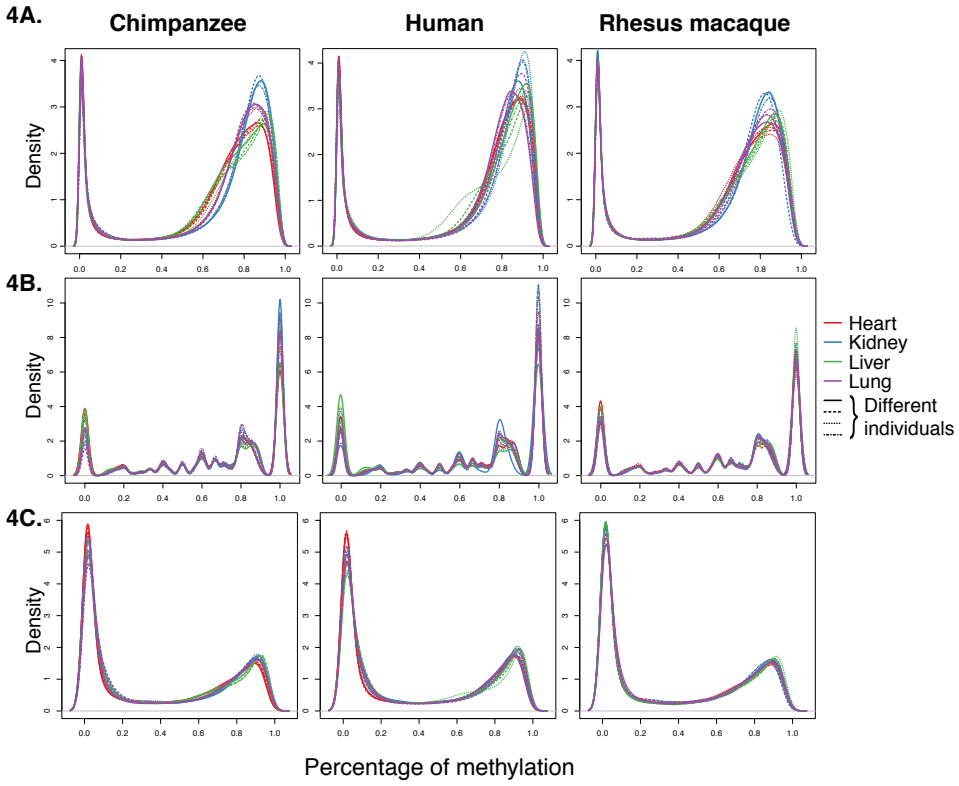


Figure 2.8: **Density function of DNA methylation levels across all species and tissues.** (A) Using raw methylation estimates at the subset of orthologous CpG sites showing a read coverage of at least 5 and no more than 10 in each sample. (B) Using smoothed methylation estimates at all orthologous CpG sites across the three species. (C) Using orthologous CpG sites located in orthologous exons, and excluding sites with C to T SNPs in any of the samples.

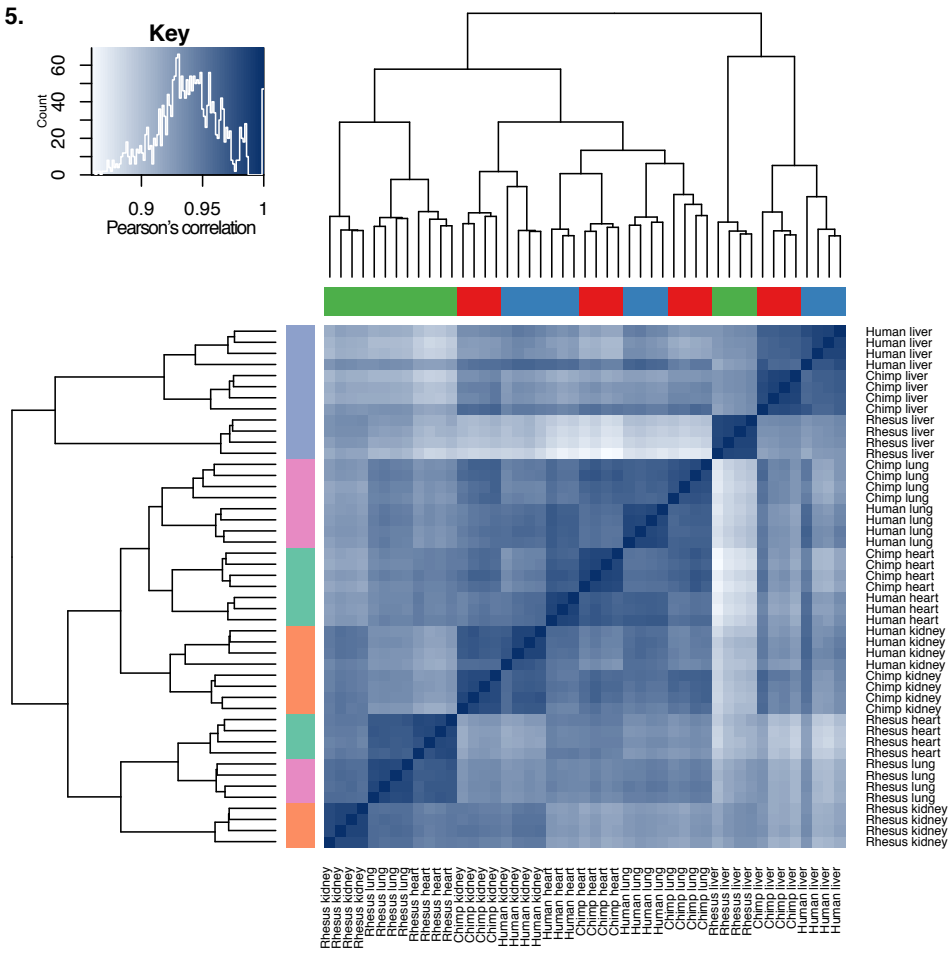


Figure 2.9: **Correlation matrix of smoothed DNA methylation levels from all orthologous CpGs.**

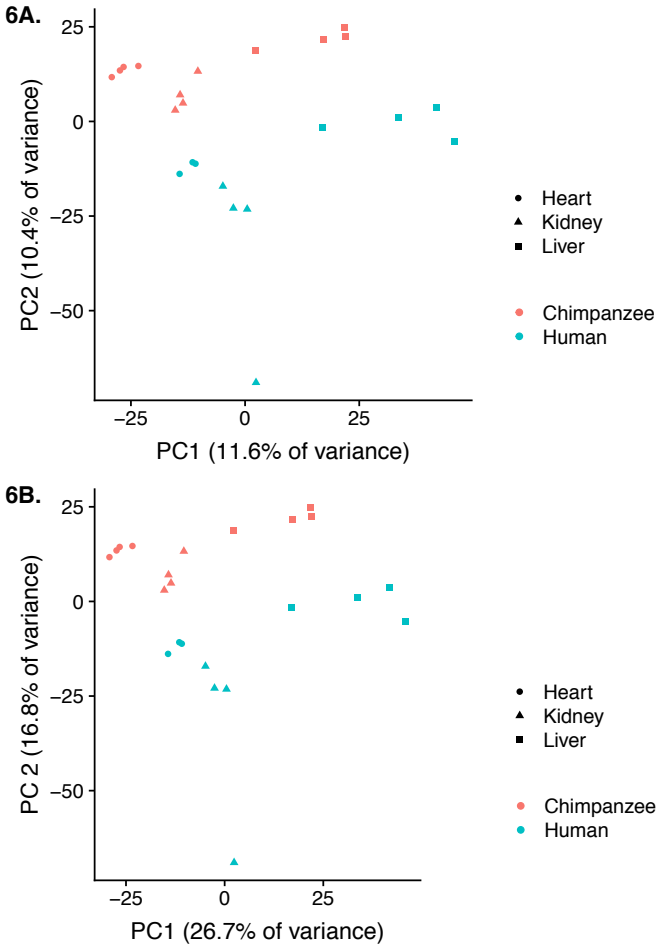


Figure 2.10: **Principal components analysis (PCA) in humans and chimpanzee hearts, kidneys, and livers.** (A) Average promoter DNA methylation values. (B) Gene expression levels in the same genes from (A).

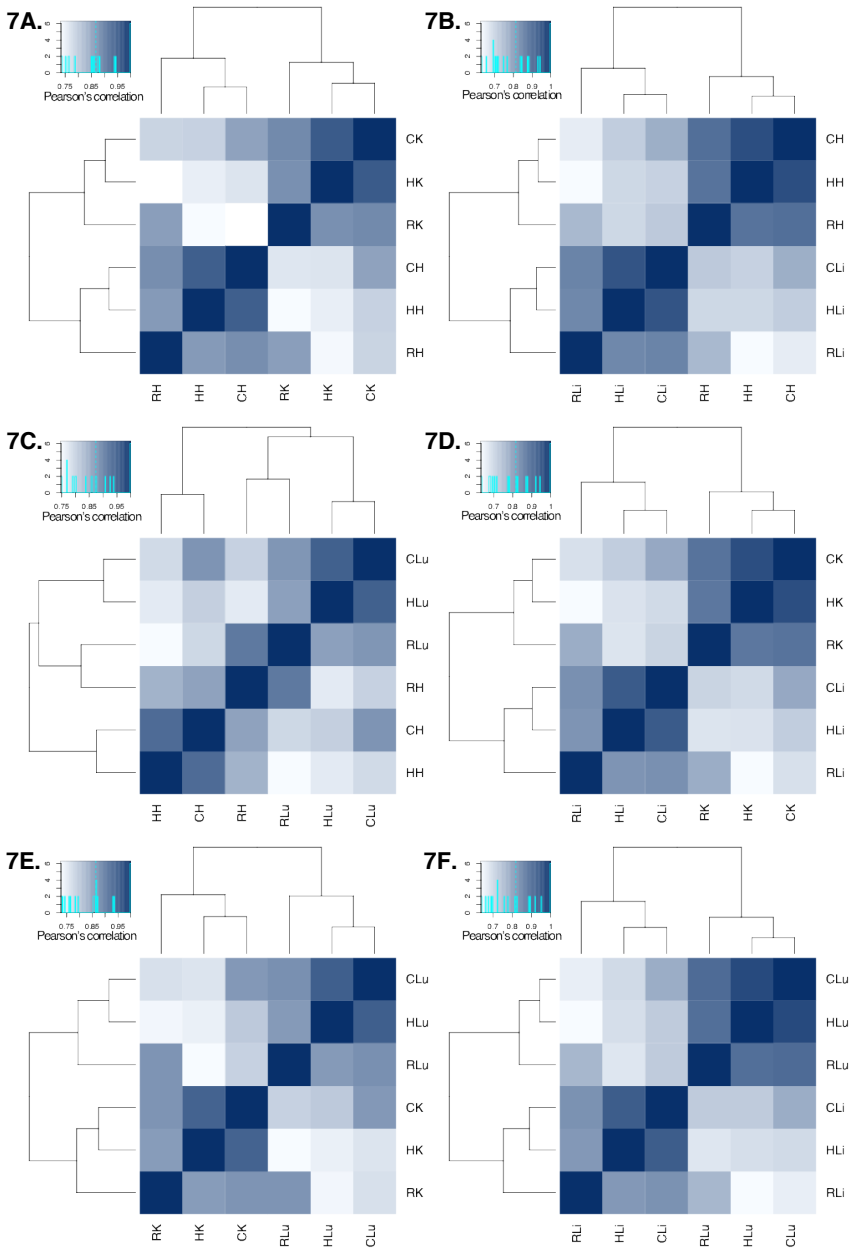


Figure 2.11: **When comparing the methylation levels of human T-DMRs and of orthologous regions in the same tissues, clustering is more highly correlated with tissue than species.** Clustering based on human (A) heart-kidney T-DMRs, (B) heart-liver T-DMRs, (C) heart-lung T-DMRs, (D) kidney-liver T-DMRs, (E) kidney-lung T-DMRs, and (F) liver-lung T-DMRs.

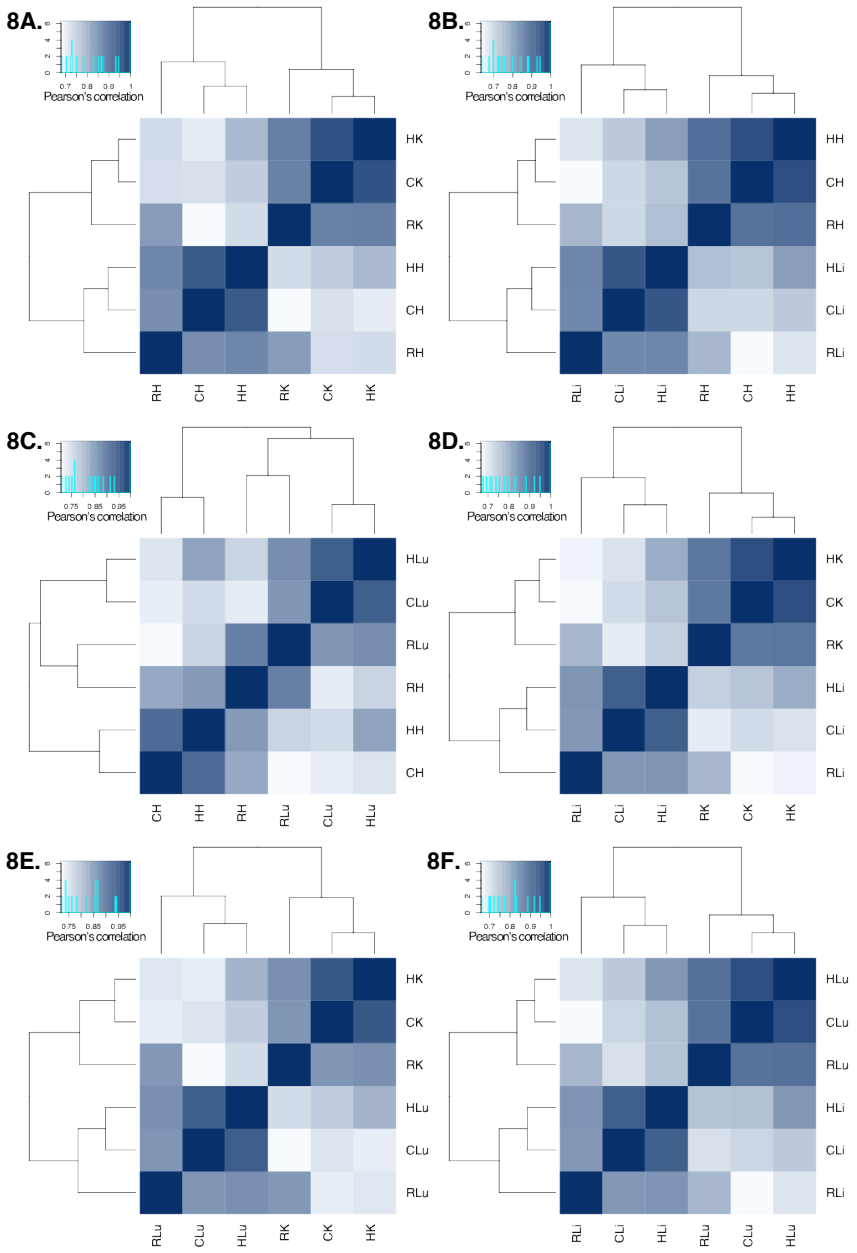


Figure 2.12: **When comparing the methylation levels of chimpanzee T-DMRs and of orthologous regions in the same tissues, clustering is more highly correlated with tissue than species.** Clustering based on chimpanzee (A) heart-kidney T-DMRs, (B) heart-liver T-DMRs, (C) heart-lung T-DMRs, (D) kidney-liver T-DMRs, (E) kidney-lung T-DMRs, and (F) liver-lung T-DMRs.

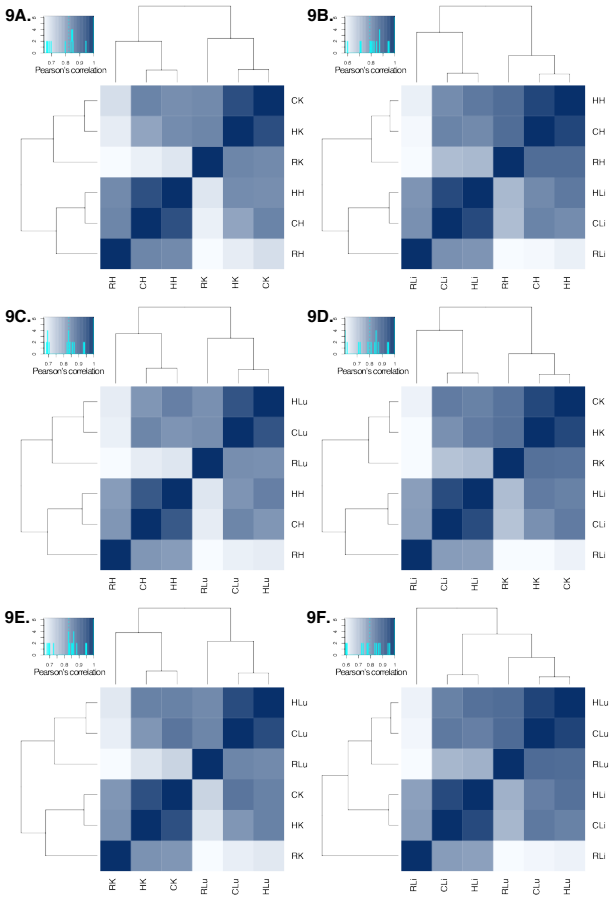


Figure 2.13: **When comparing the methylation levels of rhesus macaque T-DMRs and of orthologous regions in the same tissues, clustering is more highly correlated with tissue than species.** Clustering based on rhesus macaques (A) heart-kidney T-DMRs, (B) heart-liver T-DMRs, (C) heart-lung T-DMRs, (D) kidney-liver T-DMRs, (E) kidney-lung T-DMRs, and (F) liver-lung T-DMRs.

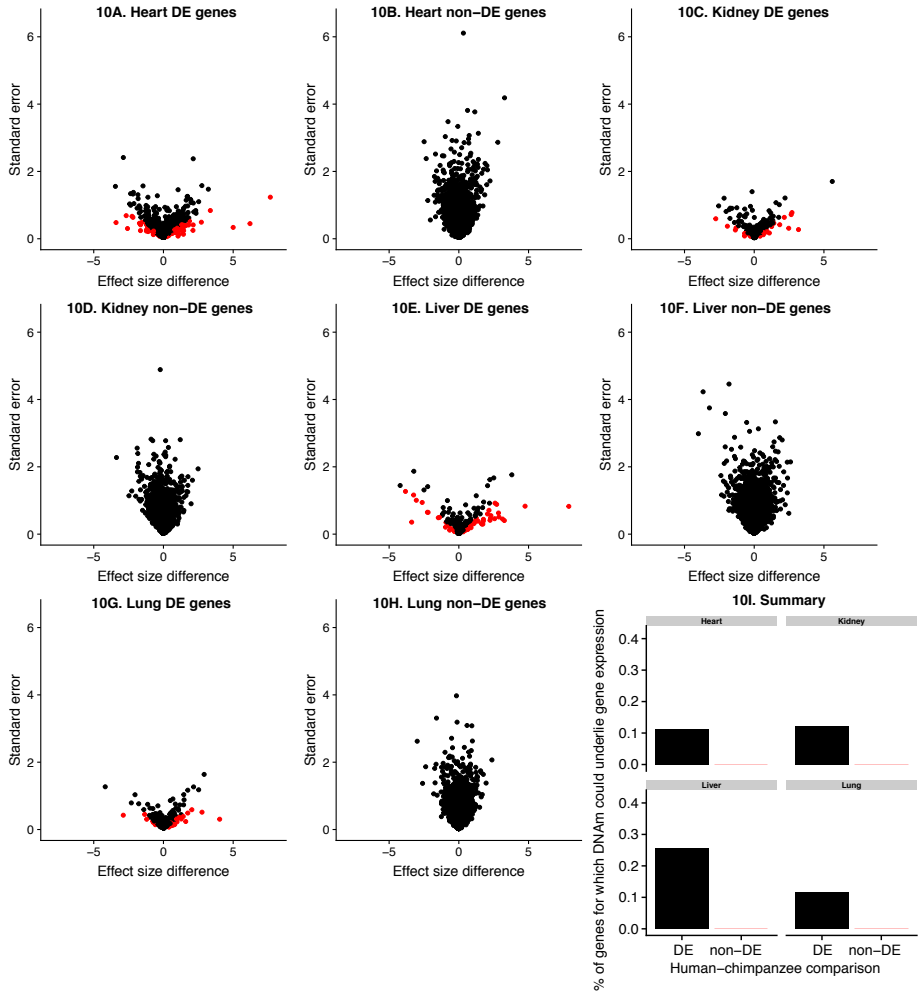


Figure 2.14: **Interspecies DNA methylation and gene expression levels (FDR = 0.05 and FSR = 0.05), in humans and chimpanzees.** Difference in species effect size before and after accounting for DNA methylation levels in genes (A) DE and (B) non-DE in the human and chimpanzee heart, (C) DE and (D) non-DE in the human and chimpanzee kidney, (E) DE and (F) non-DE in the human and chimpanzee liver, (G) DE and (H) non-DE in the human and chimpanzee lung. (I) The percentage of genes for which the evidence for inter-species differences in gene expression levels is reduced after correcting for DNA methylation levels in humans and chimpanzees.

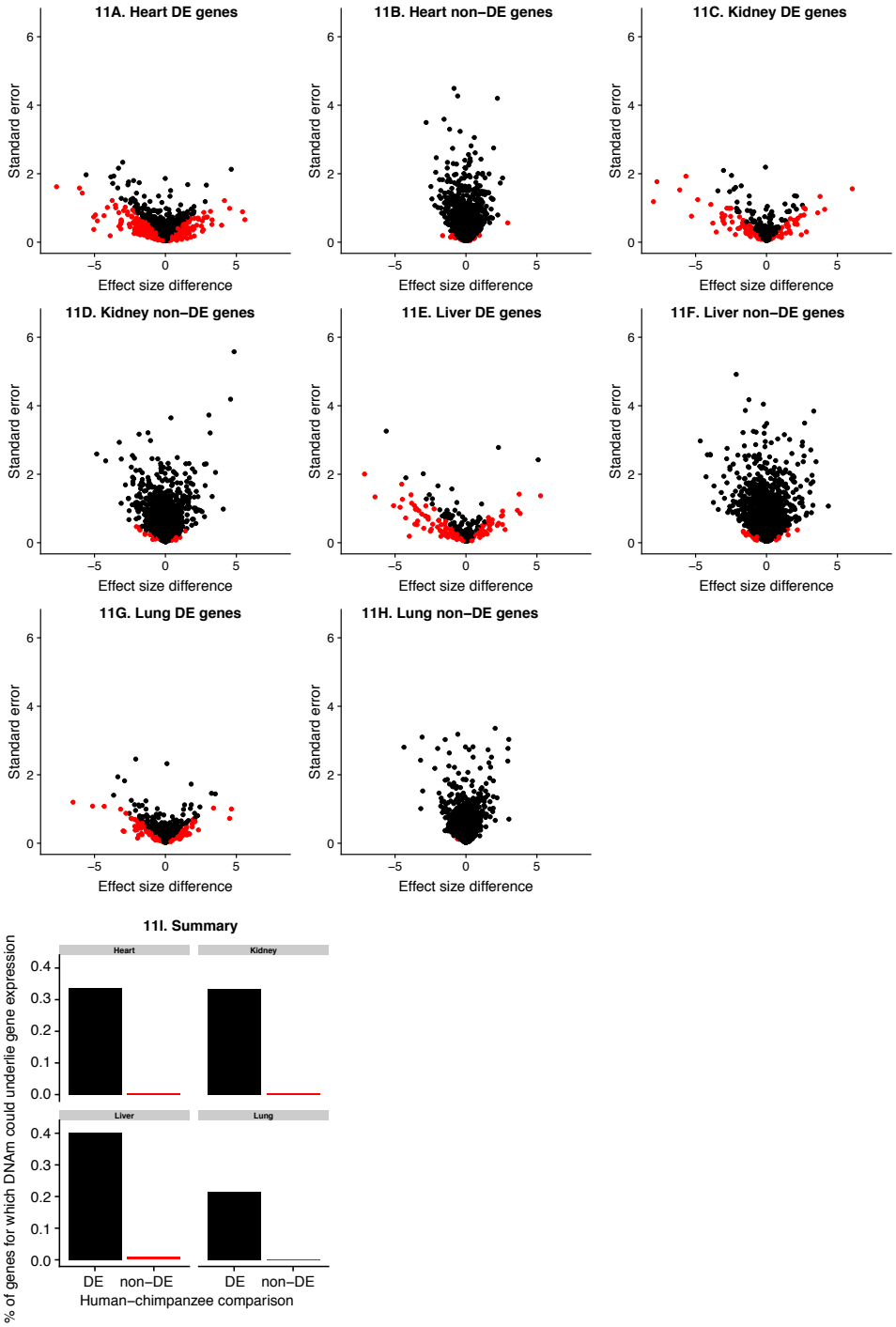


Figure 2.15: **Interspecies DNA methylation and gene expression levels (FDR = 0.05 and FSR = 0.05), in humans and rhesus macaques.** The same analysis as Figure 2.14, except in humans and rhesus macaques, rather than humans and chimpanzees.

2.9 Supplemental Tables

Table 2.2: **Recorded variables of interest.** Variables of interest related to (A) samples, (B) RNA and (C-E) DNA methylation preparation, processing and sequencing.

Table 2.3: **Normalized gene expression levels.** TMM- and cyclic loess-normalized log2 counts per million (CPM) values for 12,184 orthologous genes to be used in downstream analyses.

Table 2.4: **Technical factors and matched individuals analysis.** A-B. Results from gene expression levels. (A) Benjamini-Hochberg adjusted P values from from regression models with gene expression level PCs 1-5 as response variables and recorded biological and technical variables as explanatory variables. (B) Benjamini-Hochberg adjusted P values from regression models with species and tissue as the response variables and technical factors of interest as the explanatory variables. C-D. Results from global smoothed DNA methylation levels. (C) Benjamini-Hochberg adjusted P values from from regression models with DNA methylation level PCs 1-5 as response variables and recorded biological and technical variables as explanatory variables. (D) Benjamini-Hochberg adjusted P values from regression models with species and tissue as the response variables and technical factors of interest as the explanatory variables. E-F. DE in matched versus unmatched donors. (E) Individuals and genes. (F) Number of true versus false positive genes in subset data.

Table 2.5: **Pairwise inter-tissue and interspecies differential expression statistics.** (A) Differential expression statistics from limma and ash for each pairwise comparison. This includes the Gene name, ENSG gene ID (Gene), log2 fold change (logFC), average expression level (AveExpr), t-statistic (t), P value (P.value), q-value (adj.P.Val), log-odds (B), an estimate of standard error before adaptive shrinkage (sebetahat), local false sign rate (lfsr), local false discovery rate using estimates from adaptive shrinkage (lfdr), q-value based on the values estimated by adaptive shrinkage (qvalue), s-value based on the values estimated by the adaptive shrinkage (svalue), an adaptive shrinkage-based estimate of beta (beta_est), and an adaptive shrinkage-based estimate of the standard error (se_est) in each pairwise comparison. (B) Number of pairwise inter-tissue and inter-species DE genes at different FDR and false sign rate (FSR) thresholds. FSR is calculated using an adaptive shrinkage method. (C) Number of genes where at least 2 samples have log2(CPM) greater than 1.5 (out of 12,184 genes).

Table 2.6: **The most common pattern of inter-tissue gene expression differences is conservation between species is robust.** The observed overlap in the number of pairwise inter-tissue DE genes.

Table 2.7: **Pairwise tissue-specific DE genes are conserved.** The number of tissue-specific genes using a stringent definition (see Methods) across various thresholds.

Table 2.8: **Possible functional interpretations of genes.** Gene Ontology (GO) results for conserved genes that were upregulated or downregulated only in one tissue. Results in the (A) heart, (B) kidney, (C) liver, and (D) lung. Within each table, columns A-F represent enriched GO categories that are upregulated or downregulated only in one tissue at a less stringent definition and G-K are at a more stringent definition. (E) Average variance in interindividual gene expression levels. (F) Results from transcription co-transcription networks. (G) Results from overlap with PPI.

Table 2.9: **Group-by-tissue interactions found in Great Apes and in humans only.** (A) Number of interactions across FDR and FSR thresholds. (B) GO enrichments.

Table 2.10: **Summary information for pairwise T-DMRs and S-DMRs.** (A) Number of pairwise T-DMRs and S-DMRs (t-statistic greater than 4.6 or less than negative 4.6). (B) Overlap of T-DMRs with annotated regions, relative to 100 sets of random regions of the same length and CpG density as the T-DMRs.

Table 2.11: **Pairwise T-DMRs and tissue-specific DMRs are conserved.** (A) Conserved pairwise T-DMRs and (B) conserved tissue-specific DMRs.(C) Percentage overlap with H3K27Ac. (D) Gene Ontology results.

Table 2.12: **Percentage of genes in which gene expression differences might, at least in part, be explained by DNA methylation levels.** A-B. Percentages for interspecies DE and non-DE genes in (A) humans and chimpanzees and (B) humans and rhesus macaques. C-E. Percentages for inter-tissue conserved DE, non-conserved DE, and non-DE genes in (C) humans, (D) chimpanzees and (E) rhesus macaques. (F) Gene Ontology results.

CHAPTER 3

A COMPARATIVE STUDY OF ENDODERM DIFFERENTIATION IN HUMANS AND CHIMPANZEES

3.1 Abstract[1]

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 characterize gene regulatory changes through a four-day timecourse differentiation of iPSCs into primary streak, endoderm progenitors, and definitive endoderm. As might be expected, we find that differentiation stage is the major driver of variation in gene expression levels, followed by species. We identify thousands of differentially expressed genes between humans and chimpanzees in each differentiation stage. Yet, when we consider gene-specific dynamic regulatory trajectories throughout the timecourse, we find that at least 75% of genes, including nearly all known endoderm developmental markers, have similar trajectories in the two species. Interestingly, we observe 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.

1. Citation for chapter: Blake LE*, Thomas SM*, Blischak JD, Hsiao CJ, Chavarria C, Myrthil C, Gilad Y, Pavlovic BJ. 2017. A comparative study of endoderm differentiation in humans and chimpanzees. Genome Biology 19(1):162. * denotes equal contribution.

3.2 Introduction

Differences in gene regulation between humans and other primates likely underlie the molecular basis for many human-specific traits [118]. 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 [151, 219]. 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 [183]. 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 have resulted in important insights into the evolution of gene expression levels and the traits they are associated with [167]. Yet, we are also finding that comparative studies of gene expression patterns alone– without additional context or perturbation– provide relatively little insight into adaptive phenotypes. To a large degree, the challenge is that comparative studies in primates are extremely restricted because we only have access to a few types of cell lines and to a limited collection of frozen tissues [167]. 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 [23, 167]. 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. 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 [128]. 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 [158] and cortex development [144, 152]. 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 [227]. 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, it 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. These structures include 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 [227]. 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 [50, 81]. We differentiated the iPSCs into definitive endoderm, a process that was completed over 3 days [128], and included replicates of cell lines that were independently differentiated (see Methods; Figure 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 (Figure 3.1A). We confirmed that RNA from all samples was of high quality (Table 3.1; Figure 3.8) 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 Tables 3.2A-B. 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 [116]. 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 [23, 25]. We eliminated genes that were lowly expressed in either species, removed data from one clear outlier sample (H1B at Day 0; Figure 3.9A), and normalized the read counts (see Methods; Figures 3.9B-C) to obtain TMM- and cyclic loess-normalized log2 counts per million (CPM) values for 10,304 orthologous genes (Figure 3.2A). 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 canonical markers. These markers correspond to the cell types we expected in the different stages of differentiation (Figure 3.1B; Figures 3.10A-B). 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; Figures 3.1B, 3.10, 3.11; Table 3.3). In days 2 and 3, however, purity levels were considerably lower in the human than in the chimpanzee samples (Figures 3.1B, 3.10, 3.11; Table 3.3). 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.

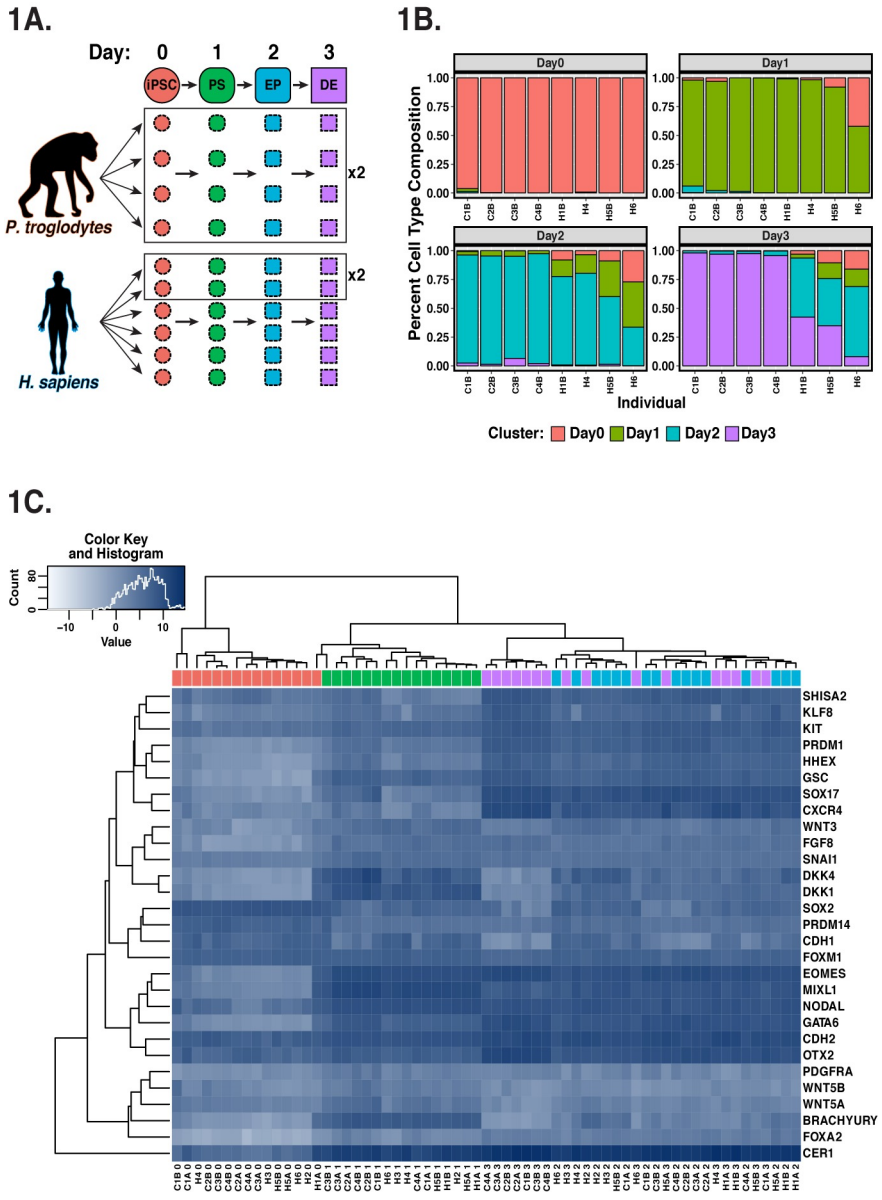


Figure 3.1: **Study design and quality control analyses.** (A) Study design. Samples from four chimpanzees and six humans were studied at four time points during endoderm development. We included two technical replicates from each of the chimpanzees and two technical replicates for two of the six humans. 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. (C) Heat map of normalized log2(CPM) as a measure of expression levels of TFs that are known to be highly expressed in one or more stages in the differentiation to endoderm. Generally, samples from the same day, regardless of species, cluster together.

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 [220], 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; Table 3.4A-D; Figures 3.9C and 3.12). 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 (Figure 3.2A; Table 3.4A-E; 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 the correlation matrix of pairwise comparisons of the gene expression levels (Figure 3.13). After characterizing global gene expression patterns, we focused on the expression of specific transcription factors with known roles in developmental pathways (Figure 3.1C) and other previously known lineage specific markers [55, 128, 197]. 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, Figure 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 [128, 197], and further supporting the validity of our in vitro system (Figure 3.14A).

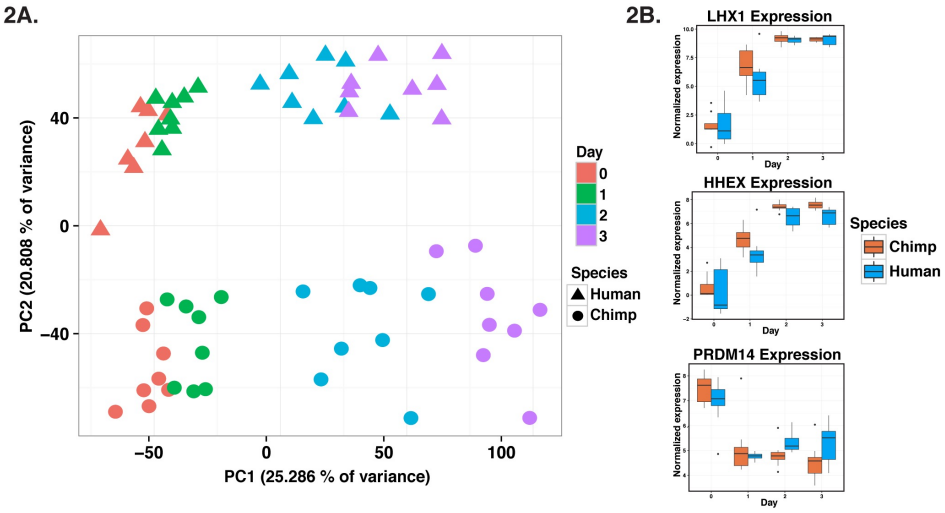


Figure 3.2: **General patterns in the data.** (A) Normalized log2(CPM) expression measurements for all genes projected onto the axes of the first two PCs. Color indicates day. Shape represents species. PC1 is highly correlated with differentiation day (r=.92). PC2 is highly correlated with species (r=0.93). (B) Box plots of normalized expression values for genes with known roles in endoderm development.

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 to 5077 genes are classified as inter-species DE at different times points; Figure 3.3A; Table 3.5A-D). 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 (Figures 3.3B-C and 3.4; Table 3.6A-F), 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 (Table 3.7). As might be expected from these observations, we found that the relationship between day and gene expression was largely independent of species (Figure 3.3D; Table 3.8A-C).

3.3.4 Joint Bayesian analysis reveals conservation of temporal gene expression profiles

In an attempt to further overcome issues of incomplete power affecting these original nave pairwise DE comparisons, and to account for dependency in data from different time points, we utilized a Bayesian clustering approach implemented by Cormotif [217]. 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 (Figure 3.5; Table 3.9A) 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, Table 3.9A; Figure 3.15A). 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 (Figure 3.5). Further, when we excluded data from

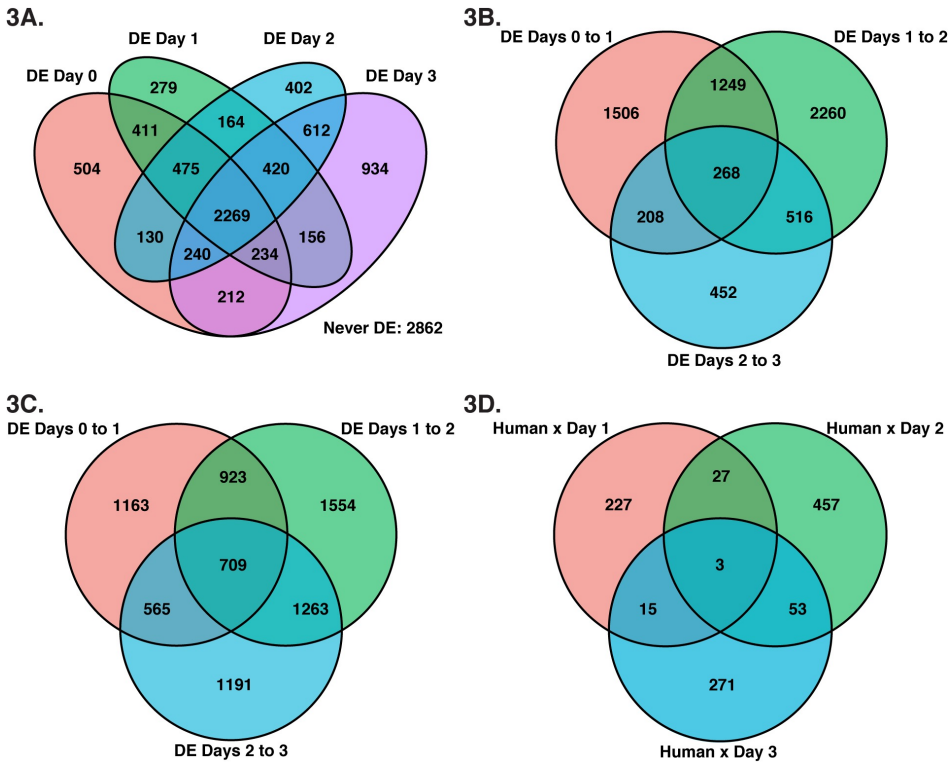


Figure 3.3: **Number of DE genes in pairwise analyses. Venn diagrams of.** (A) DE genes at each day, (B) DE genes between consecutive time points in humans, (C) DE genes between consecutive time points in chimpanzees, (D) genes with a significant species-by-time point interaction effect at each day (DE was classified at FDR of 5% in all cases).

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, the method used to combine data from technical replicates, which days were included in the pairwise comparisons, and the inclusion of all 10,304 genes in the analysis (Figures 3.15B-D and 3.16). Our observations are also robust with respect to the overall approach used to estimate and compare gene expression trajectories (Figure 3.14B). 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

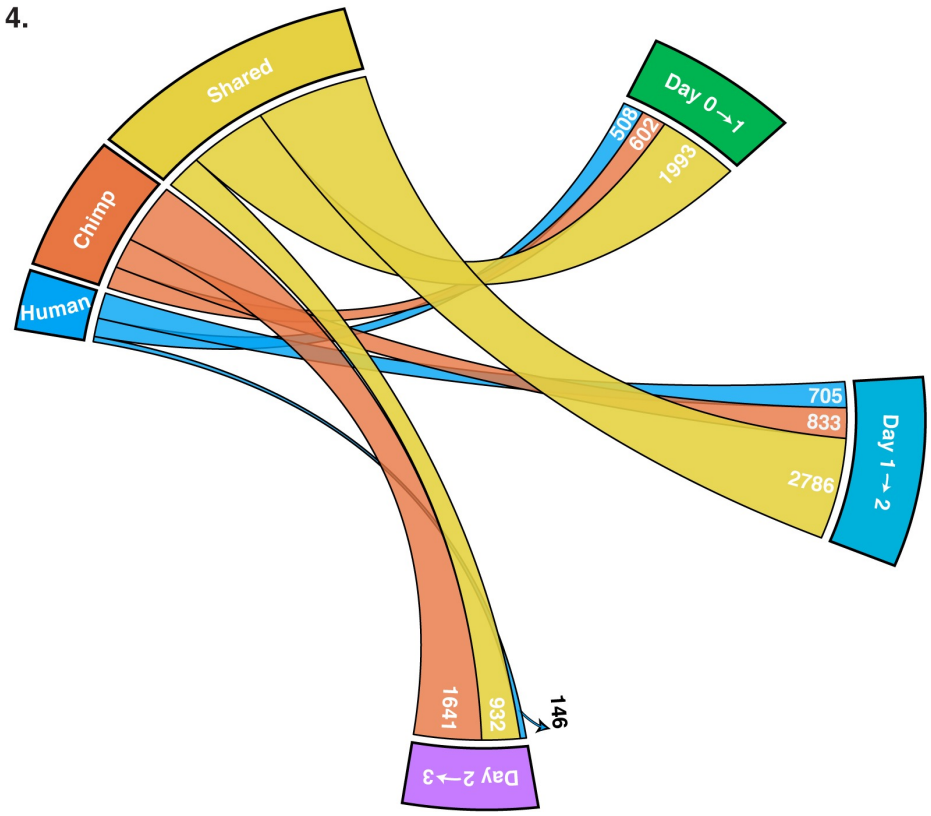


Figure 3.4: **High sharing of DE genes across species.** A circos diagram with the number of shared, human-specific, and chimpanzee-specific DE genes across time points. There is a high degree of sharing of DE genes (yellow ribbon), particularly from day 0 to 1 and day 1 to 2.

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*; Table 3.9B) [1]. These four GO annotations were not enriched in other similarly sized motifs or group of motifs (Table 3.9C-E) [1, 222]. 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.

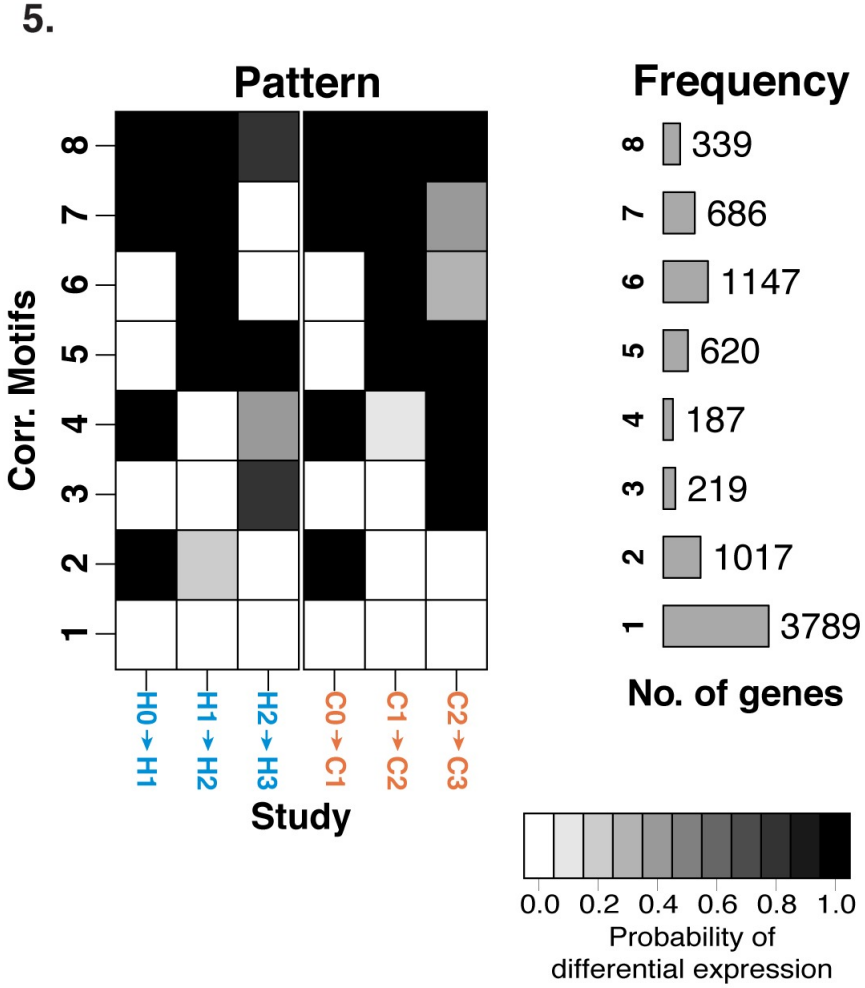


Figure 3.5: **Gene expression motifs.** Correlation motifs based on the probability of differential expression across days for each species with the number of genes assigned to each correlation motif. The shading of each box represents the posterior probability that a gene is DE between two time points in a given species. Each row ("correlation motif") represents the most prevalent expression patterns. Out of 10,304 genes, 8004 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 ([13, 87] and Figure 3.17A). 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 ($P < 10^{-15}$, Figure 3.6). We also detected this pattern when we considered the chimpanzee samples ($P < 10^{-15}$, Figure 3.6), 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 (Figure 3.17B). 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, $P = 0.04$; Figure 3.17C), 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%; Figure 3.3A; Table 3.10), we found only 279 genes that were DE exclusively in primitive streak samples (from a total of 4,408 DE genes for the primitive streak; Figure 3.3A; Table 3.10). 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; Figure 3.3A; Table 3.10). The observation of a smaller number of genes that are DE exclusively in primitive streak samples

compared with iPSCs is robust with respect to normalization method, purity of the samples, FDR cutoff, and differentiation batch (Table 3.10; Figures 3.9C, 3.17C and 3.18). 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.

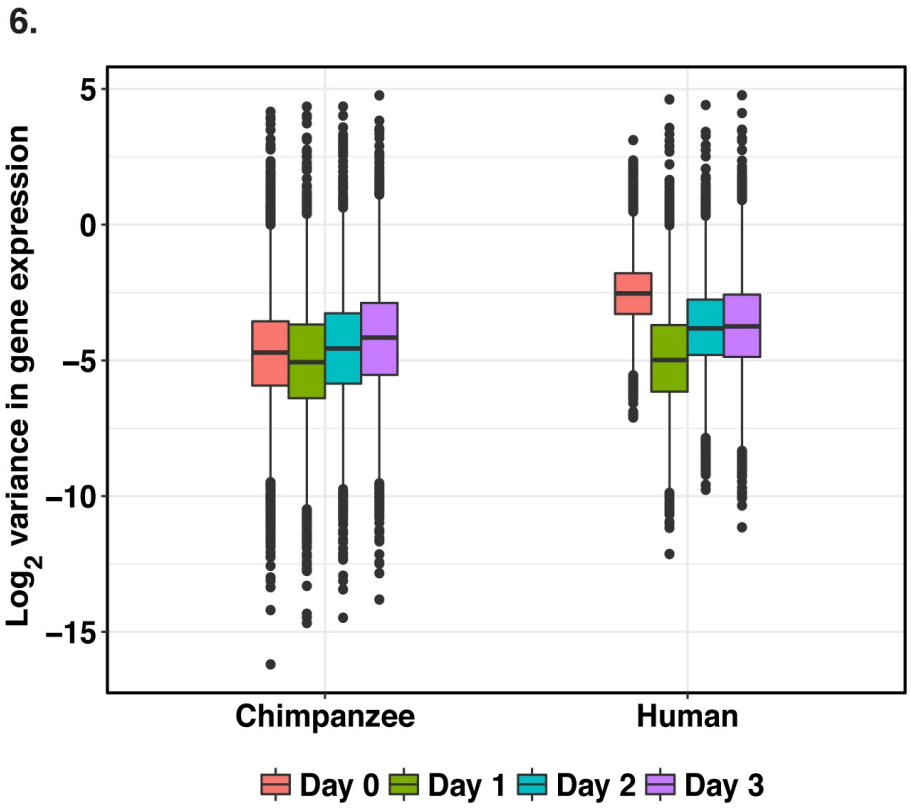


Figure 3.6: **Global reduction of variation in gene expression from the iPSCs to primitive streak state.** Box plot of the log2 variance of expression levels for each gene. Variation in gene expression levels are significantly reduced from iPSCs to primitive streak (P <10-15 in both species) but not in subsequent time points (P >0.5 in both species).

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 (Table 3.11). 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 P values from all tests can be found in Figures 3.7A-B and Table 3.12A-B, 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 P values when we tested the null hypothesis that there was no reduction in gene expression levels from day 0 to 1. Using Storey's approach [192] to account for incomplete power, we 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 (Figures 3.7A-B). This result was robust with respect to the method used to calculate the proportion of true positives [191] (Figure 3.19). We did not observe reduced variation of gene expression in any other differentiation state in our data (Figures 3.7 and 3.20A). 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 [192] 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 $<10^{-4}$, Figure 3.7C; Table 3.13). 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; Figure 3.7D). 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 (Figures 3.7E-F; Figure 3.20B). 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 (Figures 3.21-3.16). 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 [1] 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*; Table 3.14). Moreover, we found that null mutation in mouse orthologs [179] 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 $<10^{-4}$).

3.4 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 (Figure 3.5). 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 individual's 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 P value 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 [17]. 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 [211] 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 [168]. Similarly, homozygous null *MIXL1* mice have abnormal mesoendoderm development and do not survive to birth [90]. *EOMES* homozygous null mice lack trophoectoderm outgrowth [169] and do not properly form the definitive endoderm [3]. 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 [178, 211, 212]. Canalization is fundamentally linked to evolutionary states [212], 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. Our observations may be consistent with activation of deeply conserved regulatory programs at the initial stages of gastrulation followed by processes less affected by evolutionary constraint

and therefore potentially more amenable to adaptation. In other words, our results supports the expectation that gastrulation is a highly canalized and conserved process in humans and chimpanzees.More generally, we believe that despite limitations to studying comparative development using iPSC models, which we have discussed, this system provides the opportunity to study previously unappreciated aspects of primate biology.

3.5 Methods

3.5.1 Human and chimpanzee iPSC panels

In this study, we include four chimpanzee iPSC lines (2 males, 2 females) from a previously described panel [81] and six human lines (3 males, 3 females) [50] 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 [50, 81] (Figures 3.24-3.25). We identified one human individual (H5) that tested positive for episomal vector sequence (Figure 3.25). This individual was not an obvious outlier in any of our data (Figures 3.1B-C and 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 006-12. Human fibroblasts samples for generation of iPSC lines were collected under University of Chicago IRB protocol 11-0524. 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 [133]. Cell culture was conducted at 37C, 5% CO2, and atmospheric O2.

3.5.2 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 [128] 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/cm2. 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 to 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 (Table 3.2). 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 37C, 5% CO2, and atmospheric O2.

3.5.3 Purity assessment using flow cytometry

Cells were dissociated using an EDTA based cell release solution, centrifuged at 200 x g for 5 minutes at 4C 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 4C 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 4C. 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 uL 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 uL 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 [48]. Cells were stained for 1 hour at 4C 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 uL 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 [128]. Day 0 iPSC purity was estimated by dual positive OCT3/4 and SOX2 [135] as well as negative staining for EOMES. Day 1 primitive streak purity was estimated primarily based on EOMES Positive staining [71, 197] but also negative staining for SOX17. Day 2 endoderm progenitor purity was quantified by positive staining for SOX17 expression [209] (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 [48]. 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 marker's 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 Figure 3.10A, with the final purity results for the second batch of differentiation in Table 3.3. 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 (Figure 3.11, top panel) and cluster assignment on

the reduced data is shown in Figure 3.11, middle and bottom panels. We estimated cellular composition at each day by using the PCA to match clusters to a day assignment (Figure 3.1B).

3.5.4 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 (Figure 3.13B). 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 (Table 3.1; Figure 3.8). 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 (Table 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; Table 3.2A and 3.4). 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.

3.5.5 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) [116], 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 [23]. Gene expression levels were quantified using the feature counts function in SubRead 1.4.4 [123]. 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.

3.5.6 Transformation and normalization of RNA-sequencing reads

After receiving the raw gene counts, we calculated the log2-transformed counts per million (CPM) for each sample using edgeR [165]. To filter for the lowly expressed genes, we kept only genes with an expression level of log2(CPM) >1.5 in at least 16 samples per species [166]. For the remaining genes, we normalized the original read counts using the weighted trimmed mean of M-values algorithm (TMM) [166] to account for differences in the read counts at the extremes of the distribution and calculated the TMM-normalized log2-transformed CPM. When we performed principal components analysis (PCA) using the TMM-normalized log2-transformed log2(CPM) values, we found one outlier (H1B at Day 0, Figure 3.9A). 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 log2(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 [11, 164]. We found that the TMM-normalized log2(CPM) values were highly correlated with the TMM- and cyclic loess-normalized log2(CPM) values (r >0.99 in the 63 samples; Figure 3.9B). We used the TMM- and cyclic loess-normalized log2(CPM) expression values in all downstream

analysis unless otherwise stated. We calculated normalized log2-transformed RPKM values by using the function rpkm with normalized library sizes from the package edgeR [165] (Figure 3.9C). We measured the "gene lengths" as the sum of the lengths of the orthologous exons and were also used in [81]. This method of calculating RPKM was highly correlated with a method in which we subtracted log2(gene length in kbp) from the TMM- and cyclic loess-normalized log2(CPM) values (r >0.97).

3.5.7 Data quality and analysis of technical factors

To assess the data quality, we performed Principal Components Analysis (PCA) on the normalized log2(CPM) values from above (Figure 3.2A). Principal component (PC) 1 was highly associated with day and PC2 was highly associated with species (r >0.92 for each, Figure 3.2A; Table 3.4A-B). We sought to determine if the study's biological variables of interest were confounded with any of the study's recorded technical aspects (Table 3.4C-D). 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 be-

tween adaptor sequence and day (Chi-squared 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) (Table 3.4D; Figure 3.12). During this analysis, we observed that the purity estimates were relatively similar across days and between species until the final day (Figures 3.1B and 3.10C; Table 3.5). 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 (Table 3.12). The exact tools used to compare the two models were data-dependent (Table 3.2 and columns 1-2 in Table 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 P value 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 P value. 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 [146] and used ANOVA to obtain the likelihood ratio statistic and associated P value. We performed this process for each technical factor using data from days 0 and 1

as well as from days 0 to 3 (Table 3.12).

3.5.8 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 [180, 181]. 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 [121]. This limma and voom pipeline has previously been shown to perform well with n >3 biological replicates/condition [162, 185]. 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 (Table 3.4A-E), 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 (Table 3.5), DE between days for each species (Table 3.6), and significant day-by-species interactions for days 1-3 (Table 3.10). For each pairwise DE test, we corrected for multiple testing with the Benjamini and Hochberg false discovery rate [18] 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 (Figure 3.4), 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 [18]. 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.

3.5.9 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 log2(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 (Figure 3.16A)– PC1 was still correlated with day and PC2 was correlated with species (Table 3.4E). 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; Figure 3.16B). 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.

3.5.10 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 [217]. 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 and 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 (Table 3.15; Figure 3.15A). 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 (Table 3.15). The best model (the seed with the greatest log likelihood) with 7 correlation motifs is displayed in Figure 3.15B, and the best model with 8 correlation motifs is featured in Figure 3.5. 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 (expression patterns). We were initially conservative when assigning a gene to a specific correlation motif. Following the advice of the Cormotif authors [217], 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 (Table 3.9A). We also used this assignment criteria when using Cormotif to compare expression levels using

different combination methods (Figure 3.15C) and to compare all time points to day 0 (Figure 3.15D). For a trajectory to be defined as DE, the trajectory in humans and chimpanzees needed similar posterior probabilities of differential expression at each comparison along the trajectory. Using topGO [1], we tested for enrichment of Gene Ontology (GO) biological processes enrichment analysis on various combinations of correlation motifs (Table 3.9B-D). To test for significance, we used the same parameters as [26]. This included the use of Fisher's Exact Test, with topGO's weight01 algorithm to account for the correlation among GO categories in its graph structure. Categories with P value <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; Table 3.9B-D). To test this in a more rigorous way, we then compared the enriched categories for Motif 4 and 7 group given the genes in the two other groups (Table 3.9E) using the compareCluster function in the R package clusterProfiler [222]. We used a q value cutoff of 0.05 [18] and set the size of genes annotated by Ontology term for testing parameter between 3 and 3000.

3.5.11 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 log2 transformation to each variance value (Figure 3.6). 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 [87]. Three of the four tissues are derived from the endoderm germ layer, liver, lung, and pancreas. The heart tissue is mesoderm-derived, for comparison. We identified all of the genes that were expressed and passed GTEx's 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 log2 variance of the normalized gene expression using the same pipeline as before (Figure

3.17B). 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 log2 variance of these residuals for each gene. We shifted the log2 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 log2 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 log2 variances of the residuals and the log2 variances of the gene expression levels (Figure 3.17B). We then performed a one-sided t-test between the distribution of log2(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 Mann-Whitney U test on magnitudes of effect sizes (Figure 3.17C). We tested the null that there was no change in log2 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.

3.5.12 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 log2(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 P values for the F statistics from each test and plotted the densities using ggplot2 [218]. If a P value distribution appeared to be even

slightly skewed towards small P values, we used the R package qvalue to determine pi zero, the true proportion of null statistics from a given P value distribution [192]. Its complement, pi 1, is considered the proportion of significant tests from a P value 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 (Figures 3.7A-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 o and 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 (Figures 3.7A-B).

3.5.13 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 [192]. This method was recently implemented by Banovich et al. 2016 to determine the sharing of quantitative trait loci (QTLs) from different cell types [13, 192]. Using the P value distributions generated in the previous section, we subset the genes in species 2 conditioned on its F statistic significance in species 1 (unadjusted P value <0.05). To test for an enrichment of small P values, we used the P values from species 2 to determine o using the same process as the previous section (Figures 3.7C-F, 3.20). We then repeated this process for other P value cutoffs, including 0.01 and 0.10 (Figures 3.22-3.23). To determine robustness with respect to the number of genes considered significant in species 1, we calculated o for species 2 conditioned on 100 genes with the lowest P values in species 1. This process was repeated for the top 101 to all 10,304 genes (Figure 3.21).

3.5.14 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 1 based on conditioning, we performed permutation tests. First, we combined the unadjusted P values 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 [110] to permute the P values of species 1 and then merged this P value distribution with the P value distribution from species 2. We then determined pi zero in species 1 conditioned on its P value significance in species 2 (unadjusted P value <0.05). We repeated this process a total of 100,000 times and found the complement of the 100,000 values (Table 3.14). We defined the permuted null hypothesis as the mean 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 P value 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 Storey's approach [192]. 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 [1]. 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, [20, 69, 179]) and selected the "Mammalian Phenotype" option. With this output, 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".

3.5.15 Ethics approval and consent to participate

Human fibroblasts samples for generation of iPSC lines were collected under University of Chicago IRB protocol 11-0524. Informed consent to participate in the study was obtained from all human participants. The original chimpanzee fibroblast samples for generation of iPSC lines were obtained from Yerkes Primate Research Center of Emory University under protocol 006-12, in full accordance with IACUC protocols [81]. All experimental methods comply with the Helsinki Declaration.

3.5.16 Data availability

The datasets generated during the current study are available in NCBI's Gene Expression Omnibus [68] repository (GEO Series accession number GSE98411). All of our analyses using R and the subsequent results can be viewed at https://lauren-blake.github.io/Endoderm_TC/analysis/. The GTEx Vp6 dataset is available at the GTEx portal, https://www.gtexportal.org/home/datasets [87]. Data regarding genes associated with embryonic lethality response to perturbation is provided by the Mouse Genome Database at the Mouse Genome Informatics website, http://www.informatics.jax.org/batch.

3.5.17 Funding

LEB was supported by the National Science Foundation Graduate Research Fellowship (DGE-1144082) and by the Genetics and Gene Regulation Training Grant (T32 GM07197). BJP was supported by the Training in Emerging Multidisciplinary Approaches to Mental Health and Disease (T32MH020065). This work was funded by an NIH grant from NIGMS (MH084703). The content presented in this article is solely the responsibility of the authors and the funding bodies had no role in the design of the study and collection, analysis, interpretation of data, and in writing the manuscript.

3.6 Acknowledgments

We thank members of the Gilad, Stephens, and Pritchard labs for helpful discussions and comments on the manuscript. We also thank Irene Gallego Romero for her comments on the manuscript.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

3.6.1 Author contributions

YG and BJP conceived of the study and designed the experiments. SMT and BJP performed the experiments. MM and CC extracted the RNA and prepared the sequencing libraries. LEB, SMT, and BJP analyzed the results (with input from JDB and CJH). YG and BJP supervised the project. LEB, ST, YG, and BJP wrote the paper. All authors read and approved the final manuscript.

3.7 Supplementary Information

3.7.1 Supplementary Methods

A correlation-based method designed for short time-series gene expression studies

In order to classify gene expression trajectories across species, we used a Java program specifically designed for time course microarray data with a small number of time points called Short Time-series Expression Miner (STEM)[70, 182]. We used TMM-normalized log2(CPM) expression values averaged across technical replicates (n = 40) because it was unclear how we could both model expression with replicates and still use the program to complete analyses related to the overlapping trajectories across species. We note that the main paper used TMM- and cyclic loess- normalized log2 CPM data, but the two datasets are

highly correlated (r >0.99). Per the STEM manual's suggestion for removing gene expression trajectories driven solely by technical noise, we prefiltered our set of 10,304 orthologous genes using a minimum difference of at least one log2 fold change between the minimum and maximum values (not necessarily across consecutive days). We also filtered out the genes that were in the bottom quantile of the within-species correlations. We chose a quantile cutoff rather than a value cutoff for each species because the chimpanzee samples were generally more highly correlated than the human samples. To ensure that the same genes were used for analysis, filtering was done outside of STEM using a custom R script and the union of the gene passing filtering criteria in either species was utilized (i.e., a gene only needed to pass inclusion criteria in one of the two species to be considered for further analysis). 3940 genes remained post-filtering and were used to first run the chimpanzee and human expression data separately (1 run each). Using the Clustering Method option, we used the "STEM Clustering Method" and selected the options Maximum Number of Model Profiles = 125 and Maximum Unit Change in Model Profiles between the Time Points = 2. The former was selected so that we could still maintain a reasonable number of profiles (maximum of 125) and the number of significant model profiles was relatively robust to the maximums unit change in model profiles between time points. Since we had previously filtered the data, when we reached the filtering stage in the program, we selected values in the STEM program that would not cause any new genes to be removed. Specifically, we used the following options: Maximum Number of Missing Values = 0, Minimum Expression Change = 0, Minimum Correlation between Repeats = -1. We also normalized the data so that the first time point had a gene expression level of 0 and all subsequent values would be log2 fold changes relative to the first time point. We used the STEM package to calculate the correlation of each gene to each model profile and to assign each gene to the model profile with which it had the highest correlation. We noticed during early runs of this program using expression values from either species that there were many highly correlated model profiles which we felt would still be well represented

but be more interpretable if they were combined into a single profile. Therefore, we selected the option to collapse highly correlated model profiles into a single representative model profile. After each of our 3260 genes were assigned to model profiles, we used the package to determine the significance of the number of genes assigned to each profile by a permutation test. We therefore set options Maximum Correlation = 0.9, Number of Candidate Model Profiles = 1,000,000, Number of Permutations per Gene = All permutations and Significance Level = 0.05 with Correction Method = Bonferroni for the assignment and test of significance steps. Through this process, we acquired the lists of all model profiles with at least 1 gene assigned to it, the gene names assigned to each model profile, and which model profiles were significant in each species. We used the "overlap" option in STEM to determine which profiles were significant in both species or only significant in one species. To find genes that were significantly enriched in correlated clusters, the compare option was used within STEM using a maximum uncorrected intersection P value of 0.005 and only considering profile intersections of at least 20 genes.

3.7.2 Supplementary Text

Joint Bayesian analysis provides relatively consistent estimates of the conservation of gene expression patterns

As stated in the main text, the joint modeling technique Cormotif [217] leveraged expression information shared across time points to identify the most common temporal expression patterns. In the main paper, we estimated the degree of conservation of trajectories based on a model with 8 correlation motifs to be 75%. When we choose the most likely model with 7 correlation motifs instead of 8 and used the same process to calculate the estimate as in the main text, this percentage rose slightly to 80% (Figure 3.15B). We also calculated the degree of conservation of gene expression trajectories using all 10,304 genes. We did so by using other values generated by Cormotif, the posterior probabilities of differential

expression (DE) for every gene at each state transition in each species (e.g. the posterior probability of a given gene being DE in human samples between days 0 and 1). We calculated the percentage of genes with similar posterior probabilities of differential expression (<0.20) across species at each comparison along the trajectory (e.g. a difference of <0.20 between the posterior probabilities of DE in the human and chimpanzee samples from days 0 to 1, as well as from day 1 to 2 and from day 2 to 3). The results suggested relatively strong conservation of gene expression trajectories across species, as 67% of the 10,304 genes tested had similar posterior probabilities of DE across species at all three comparisons. This percentage was similar (71%) when we calculated this estimate with 7 correlation motifs instead of 8. Overall, our estimates of the degree of conservation using analysis from Cormotif were relatively consistent (range: 67%-80%).

The estimates of the conservation of gene expression patterns are relatively consistent when using a joint Bayesian method and a correlation-based method

The joint modeling approach Cormotif allowed us to utilize the multiple data point structure of our data. There are also some limitations to this method, however. For example, genes are grouped based on the probability that they are DE but the approach does not take into account directionality. Therefore, genes with increased and decreased expression could theoretically be included in the same correlation motif. Furthermore, Cormotif ultimately relies on pairwise comparisons, albeit among all data points. To address some of these considerations, and to ensure that our results are robust we used an additional approach to analyze the data. We utilized an approach specifically designed to group genes with similar expression patterns for time-series studies with 8 or fewer time points, called Short Time-series Expression Miner (STEM) [70]. After filtering genes that only showed small changes in expression as recommended by the program (see Supplementary Methods), 3940 genes with relatively large changes in gene expression levels over the timecourse remained for analysis. In the human samples, 3260 of these genes clustered into 9 significant

model profiles (Figure 3.14B). (There were an additional 20 profiles that did not contain enough genes to reach significance.) 2846 were clustered into 6 significant model profiles in the chimpanzees (Figure 3.14B). (There were 24 additional non-significant profiles for the chimpanzee samples.) All of the significant chimpanzee profiles are also significant profiles using human samples, highlighting the similarities of gene expression trajectories across the species. Supporting our assessment of the high degree of conservation, the 3 significant human profiles that did not reach significance using chimpanzee samples nevertheless contained the same ranking (by number of genes) in each species. To determine the proportion of genes with conserved trajectories for this method, we analyzed all genes assigned to the same cluster (1685 genes). We then identified an additional 1020 genes in similar clusters (mean profile correlation = 0.73, see Supplementary methods below). Using this approach, we discovered 2705 genes (or 69% of the 3940 genes analyzed) with the same or a highly correlated trajectory across species. This high degree of conservation across the timecourse is very similar to the estimates provided by the joint Bayesian method. We found that the STEM clustering frequently assigned known regulators or markers of the differentiation process to the same cluster. This increased our confidence in the results, as we had a prior expectation that these markers would be assigned to the same or similar profiles for both species. For example, *EOMES* and *MIXL1*, transcription factors essential for endoderm formation, are assigned to the same clusters across species (Figure 3.14A). Furthermore, of all the developmental regulators and markers examined (a total of 34 gathered from literature) the majority (70%) are assigned to the same cluster in both species, and an additional 20% assigned to highly similar profiles (correlation = 0.70). Most of these 34 genes (61%) are assigned to just two profiles, suggesting that these profiles are likely the representative trajectories for developmental progression and drivers of endoderm specification.

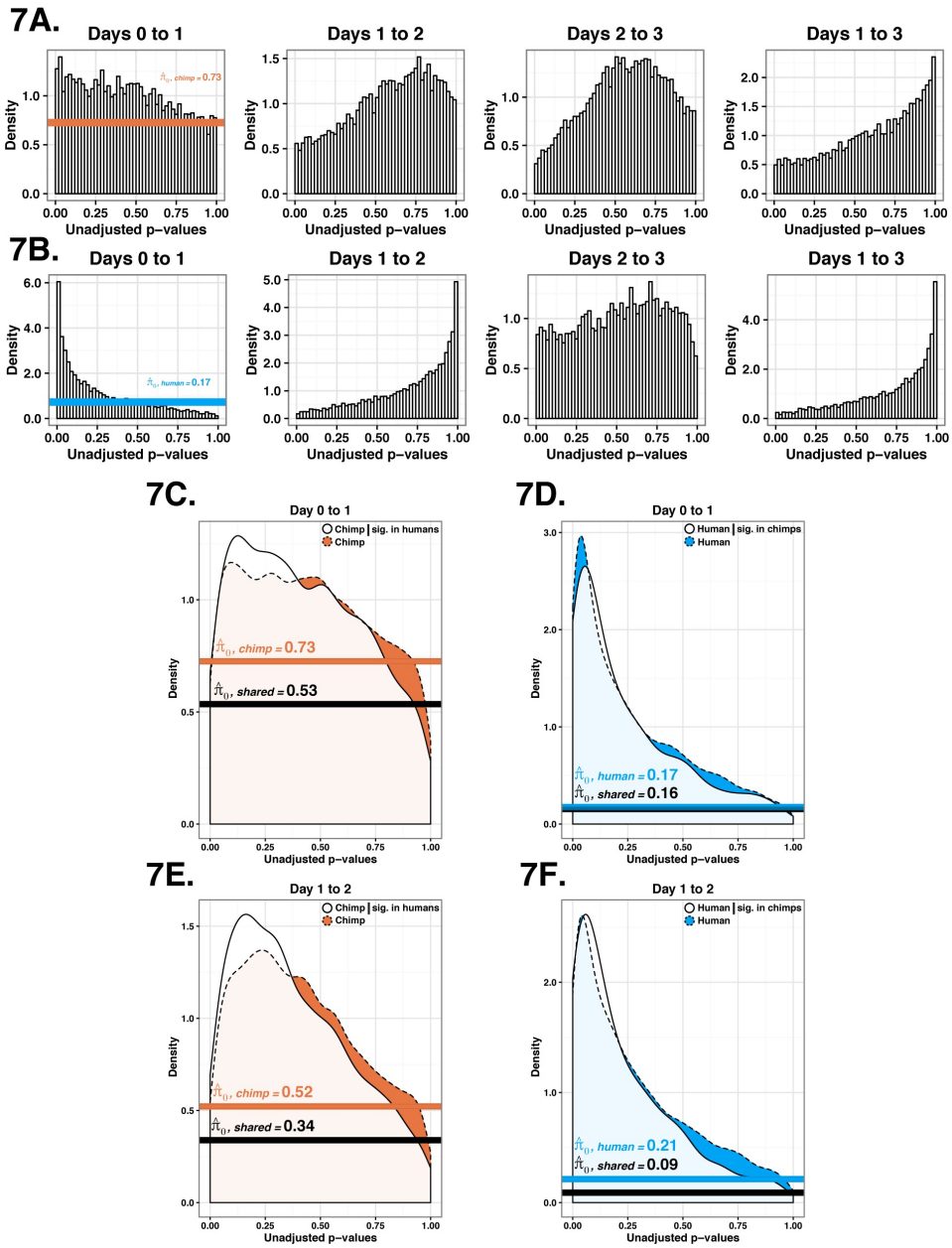


Figure 3.7: **Conserved patterns of reduced variation in gene expression at primitive streak.** We plotted the P value distributions of F tests of the null hypothesis that there is no reduction in variation in gene expression levels as samples progress along the time course in human (A) and chimpanzee (B) samples. Pi zero is the estimated proportion of null tests in each distribution. In the next four panels, we plotted the P value distribution for the same test, but included only genes whose variation was classified as reduced between states in the other species; thus in (C) we plotted P value distributions of F tests in chimpanzees only for genes who variation was classified as reduced (P <0.05) in humans and in (D) we did the reverse. In (E) (chimpanzee conditional on human) and (F) (human conditional on chimpanzee), we plotted the P value distributions of F tests of the null hypothesis that there is increase in variation in gene expression levels as the samples progress from day 1 to 2.

3.8 Supplementary Figures

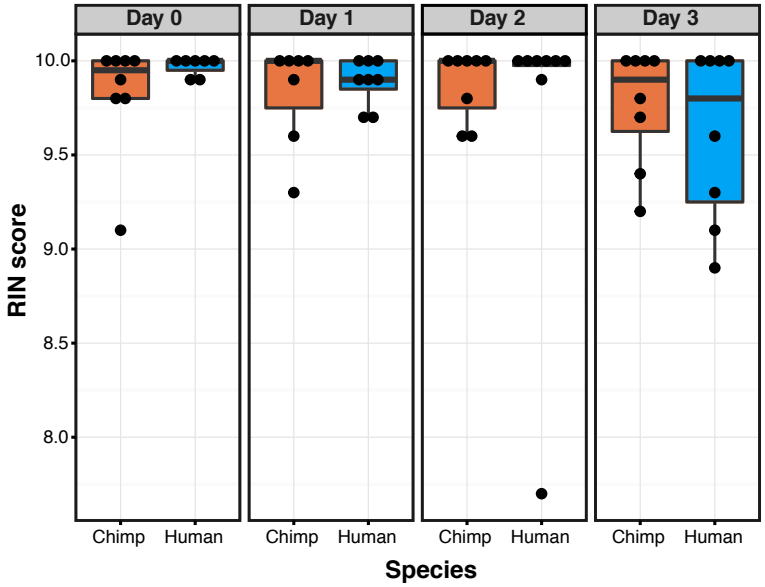


Figure 3.8: **RNA Integrity Number (RIN) scores across biological variables of interest.** RNA quality was not confounded with day, species, or batch (Benjamini-Hochberg adjusted Pvalue > 0.10 for individual linear models). Each point represents one sample.

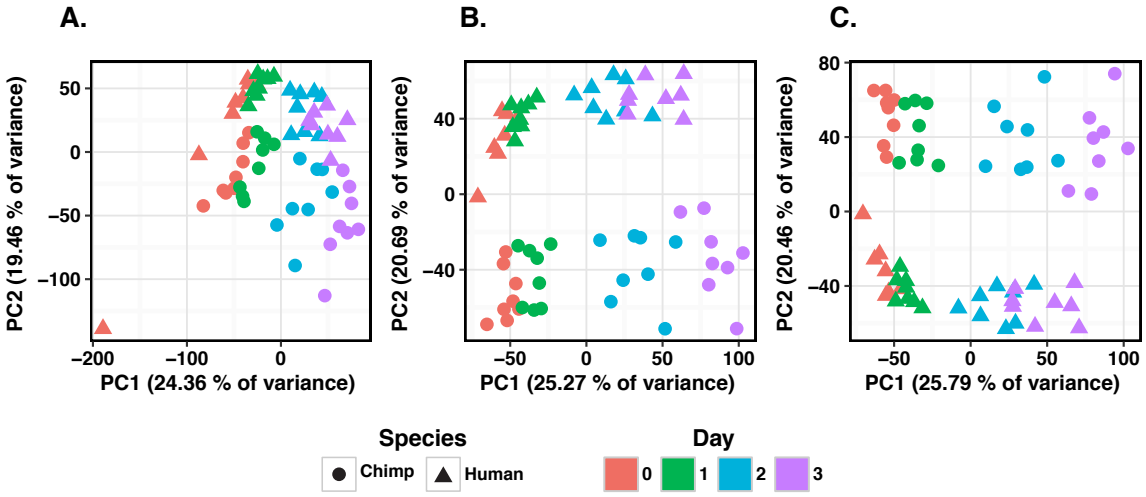
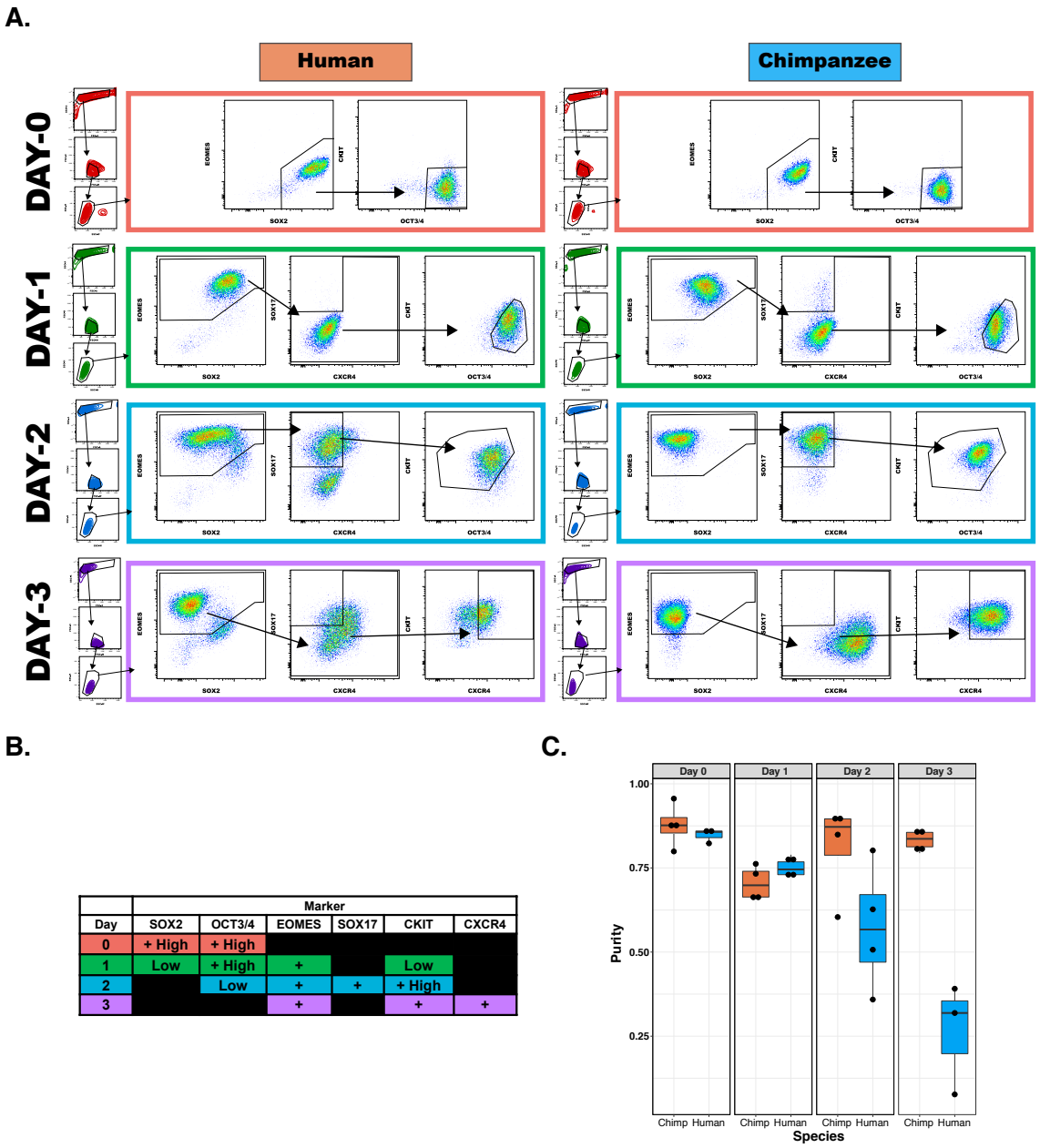


Figure 3.9: **Principal components analysis (PCA) of normalized data.** A. Prior to the removal of one outlier sample (H1B at day 0). PC1 separates the data from H1B at day 0 from all other samples. B. PCA of the TMM-normalized $\log_2$-transformed CPM. PC1 separates the samples by time point (day). 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 3.10: **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.

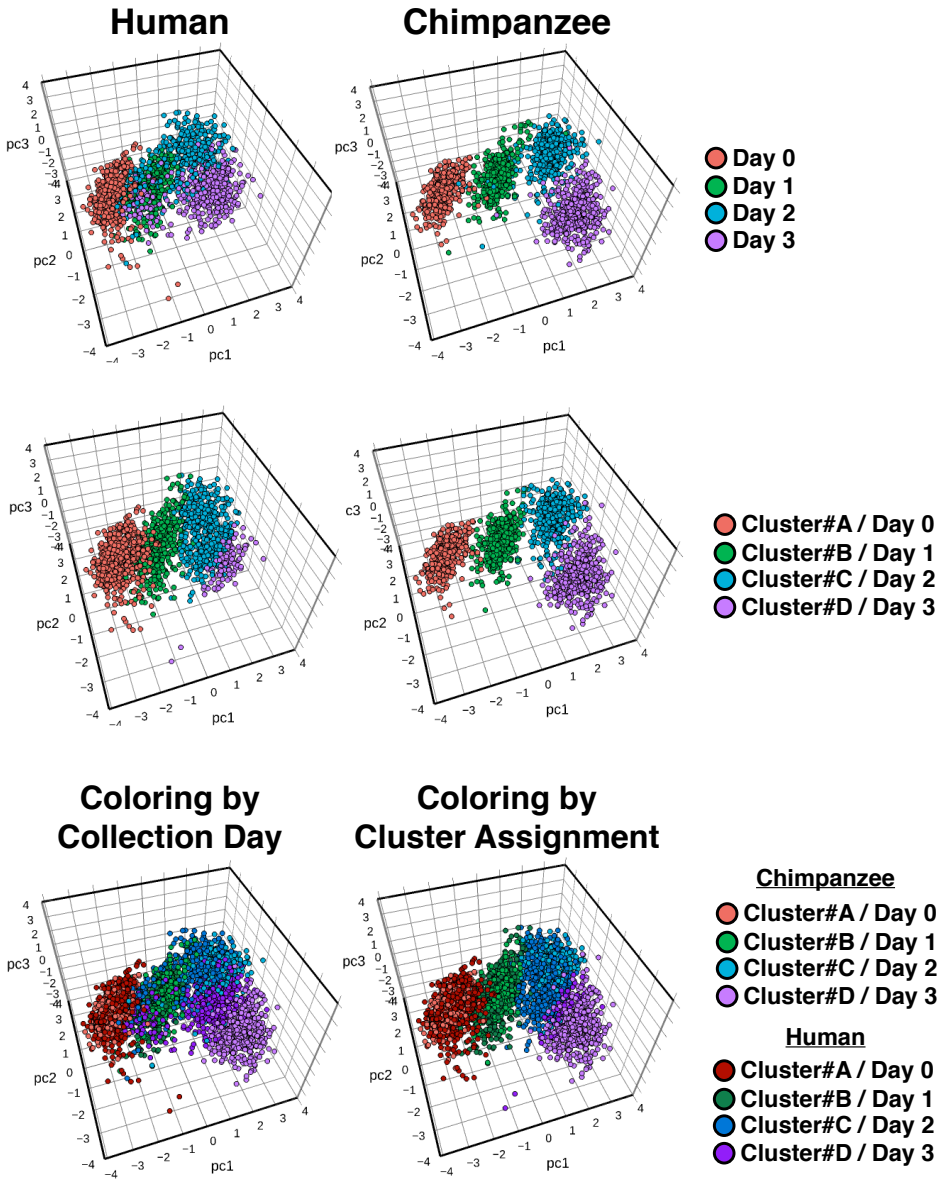


Figure 3.11: **Clustering of fluorescence values.** Compensated fluorescence values projected onto the axes of the first three principal components. For plotting, we randomly sampled 400 cells from each individual (total of 1600 cells per species) from each day separately in humans (left panel) and chimpanzee (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.

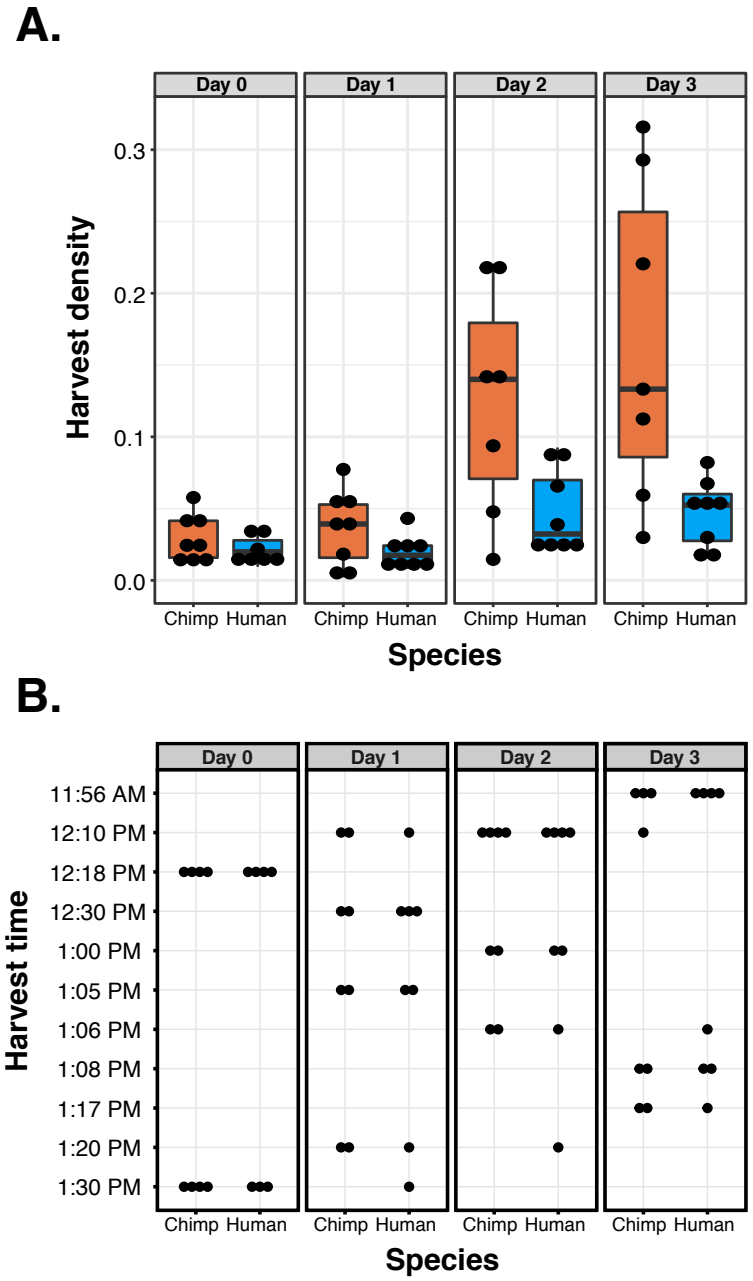


Figure 3.12: **Distribution of potential confounder variables by day and species.** A. Harvest densities for each sample by day and species. B. Harvest times for each sample by day and species. For more information about harvest densities or times.

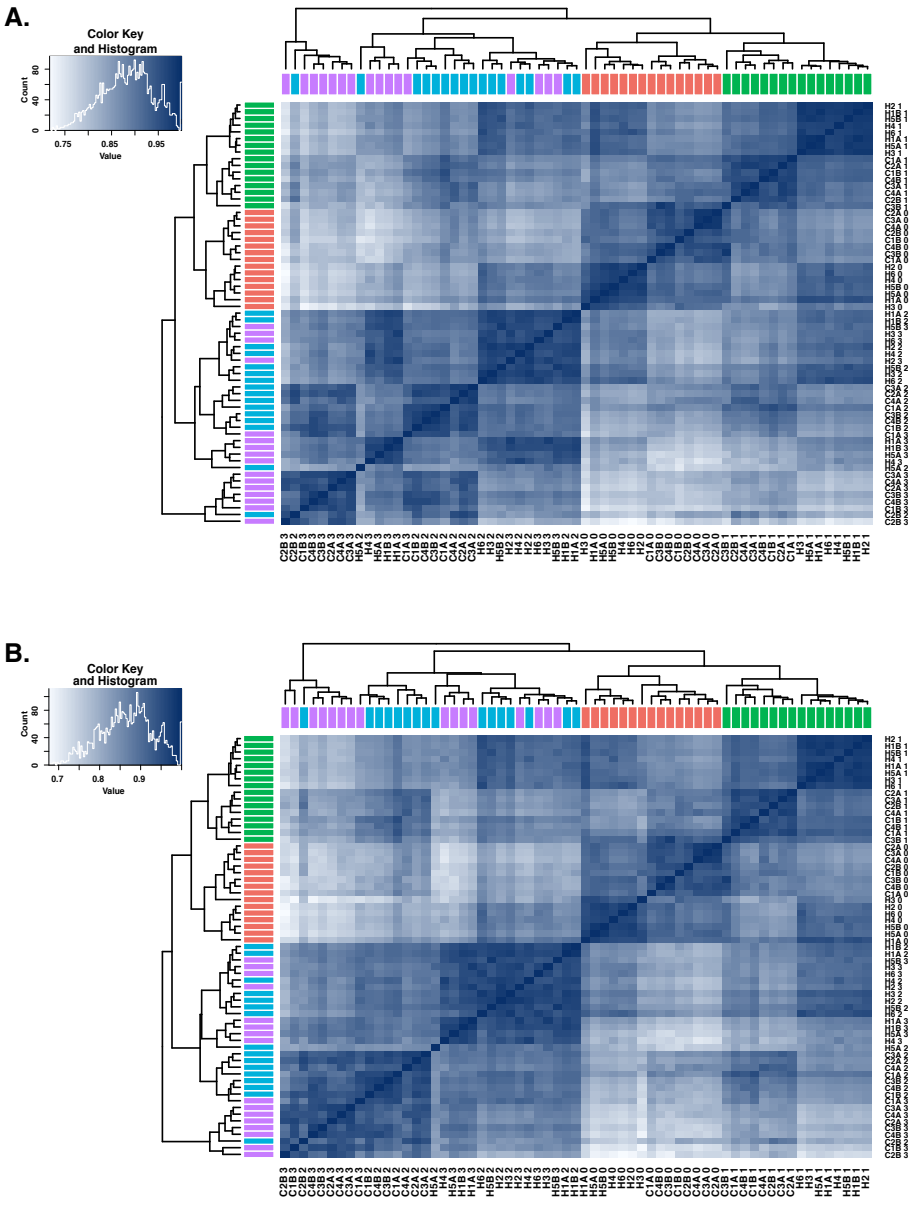


Figure 3.13: **Correlation matrices of normalized $\log_2$(CPM) gene expression values from 10,304 genes.** A. Heat map of the Spearman's correlation matrix of normalized $\log_2$(CPM) gene expression values from 10,304 genes. Each square represents the Spearman's correlation of the normalized expression 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 Pearson's correlation matrix of normalized gene expression values from 10,304 genes. The results are similar to A.

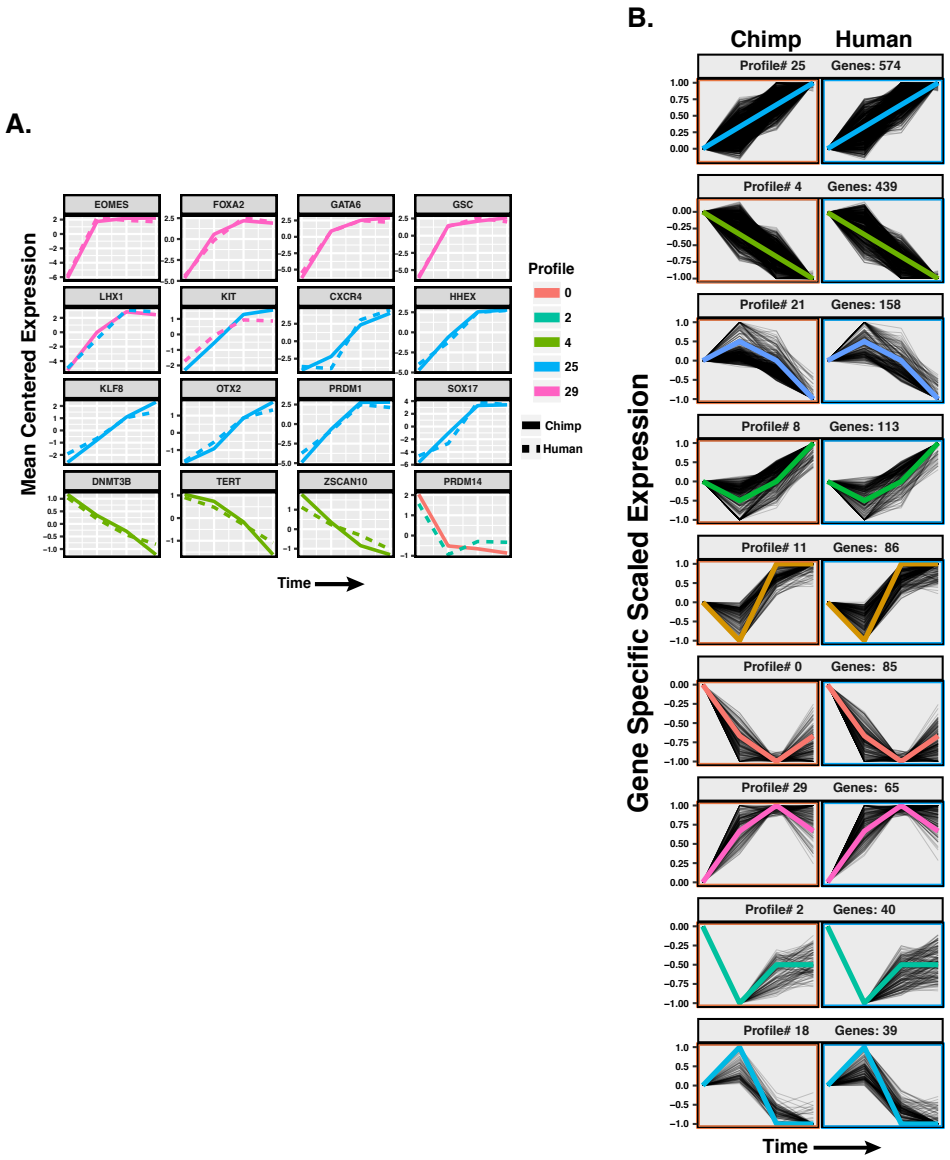


Figure 3.14: **Classifying genes into temporal profiles with Short time course expression miner (STEM) based on TMM-normalized $\log_2$(CPM) expression values.** A. Temporal profiles for 16 known critical regulators or markers of the differentiation process. Chimpanzee time course is shown using a solid 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.

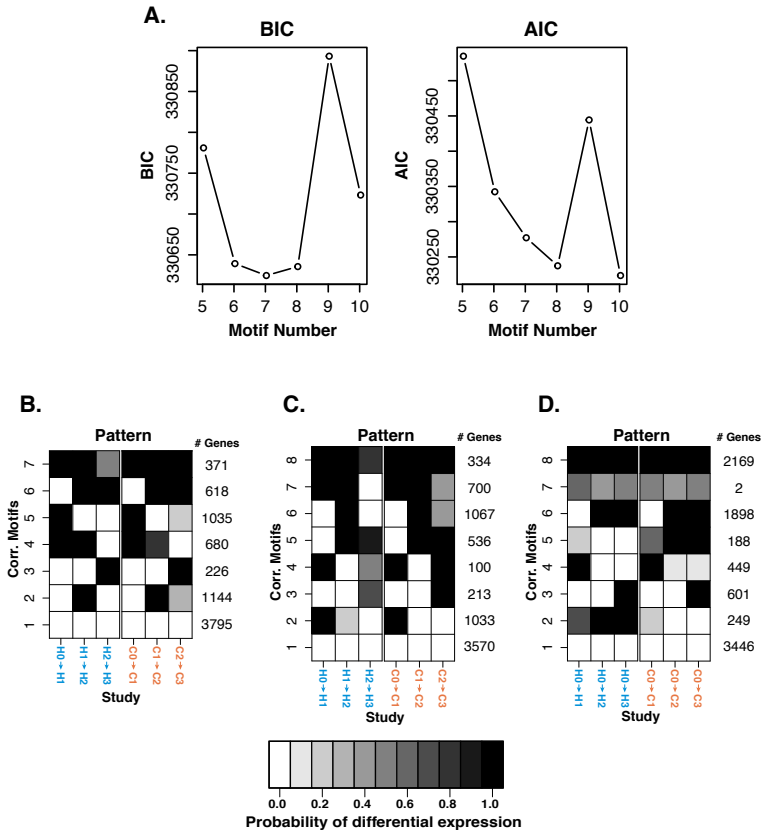


Figure 3.15: **Assessing the robustness of Cormotif results.** The shading of each box represents the posterior probability that a gene is DE between two time points in a given species. Each row ("correlation motif") represents the most prevalent expression patterns along the trajectory. The Bayesian 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 3.5 (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 3.5. 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).

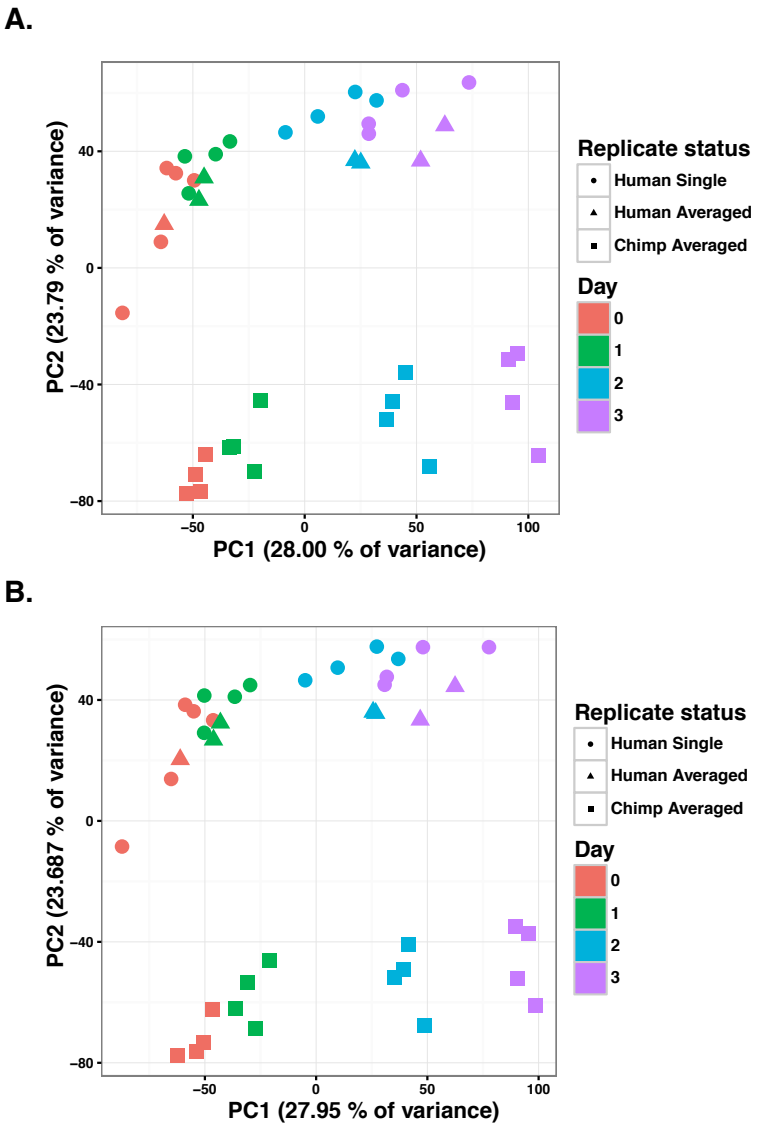


Figure 3.16: **Principal component analysis (PCA) of normalized, combined data (n = 40)** A. PCA of the normalized data after averaging the $\log_2$(CPM) gene expression values across technical replicates (n = 40 data points). PC1 is highly correlated with time point (day) and PC2 is highly correlated 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.

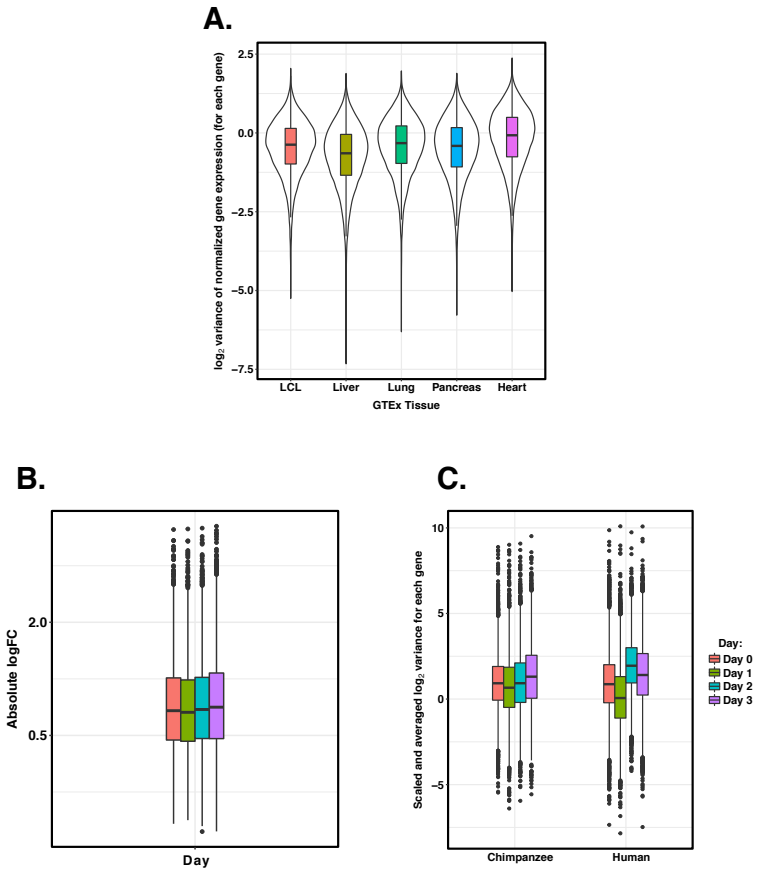


Figure 3.17: **Assessing the variance of gene expression levels.** A. Boxplot of the $\log_2$ variances of normalized RPKM gene expression levels from lymphoblastoid cell lines (LCLs) and 4 tissues collected by the GTEx Consortium. The distribution of variance of gene expression levels in LCLs are similar 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.

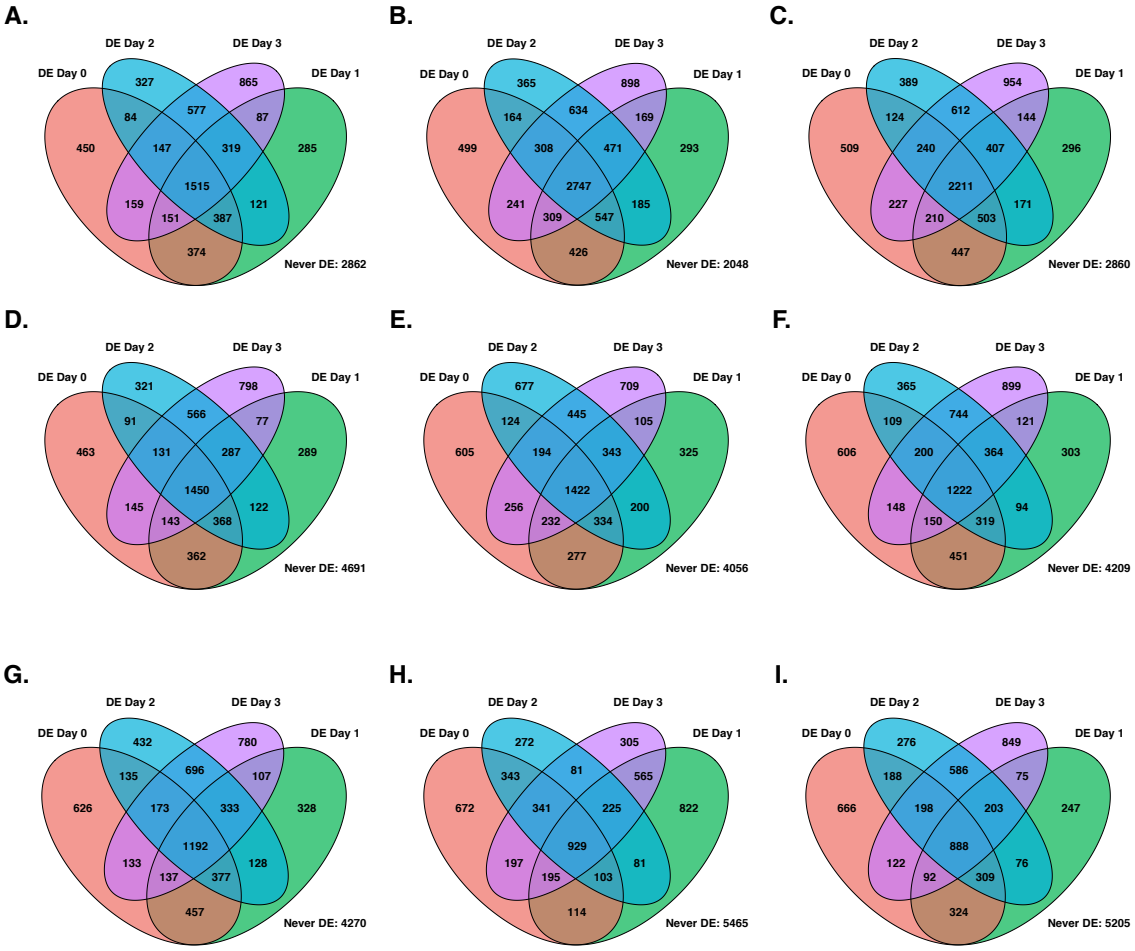


Figure 3.18: **The number of interspecies DE genes using a pairwise, linear model-based approach (limma) is lowest in the primitive streak state.** A. Venn diagram of all DE genes across species at 1% FDR. B. Venn diagram of all DE genes across species at 10% FDR. C. Venn diagram of all DE genes across 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.

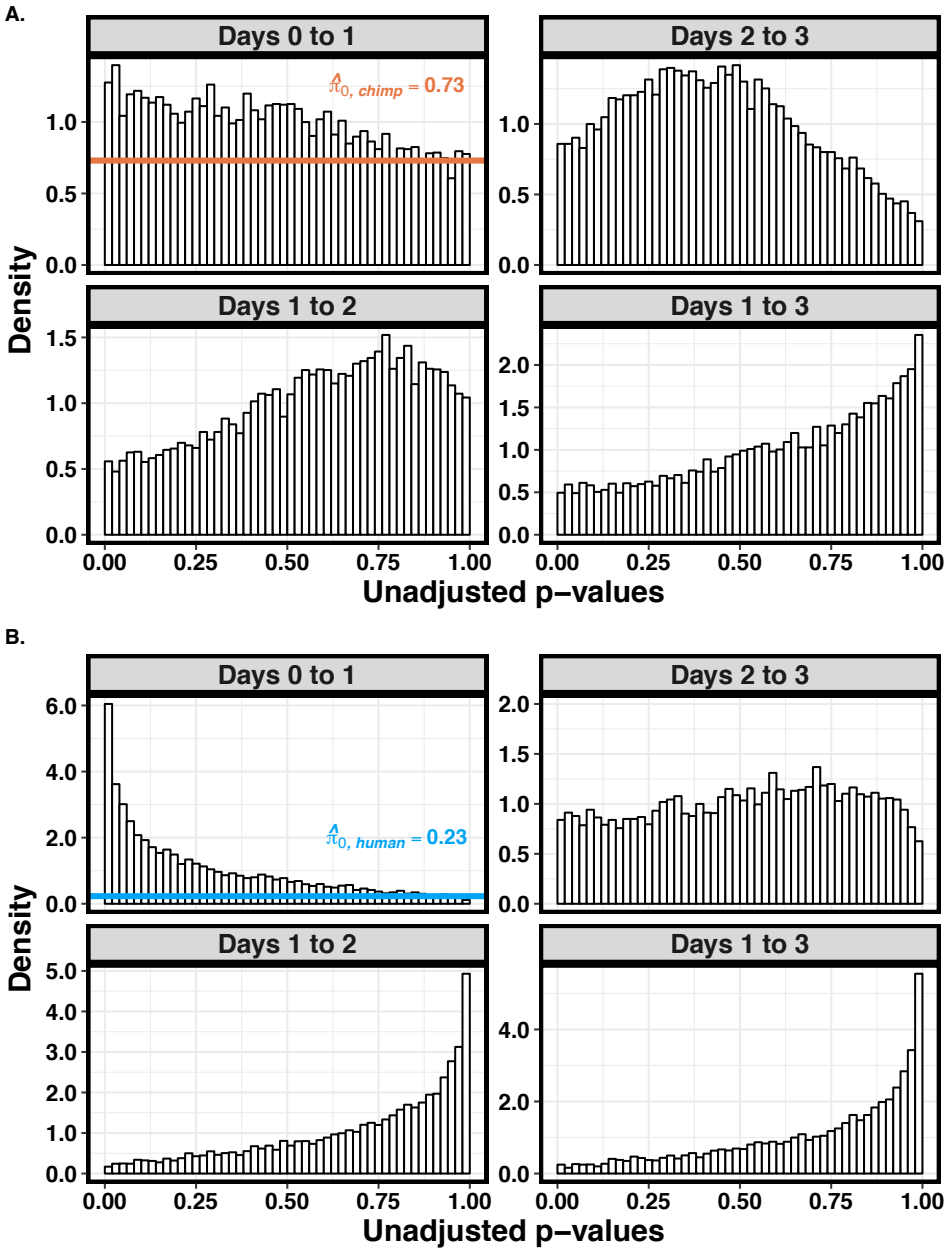


Figure 3.19: **Reduced variation in gene expression at the primitive streak is maintained when using a bootstrap method to calculate pi zero.** Results from F tests against the null hypothesis that there was no reduction in variation in gene expression levels (day 1 versus 0, 2 versus 1, 3 versus 2, 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.

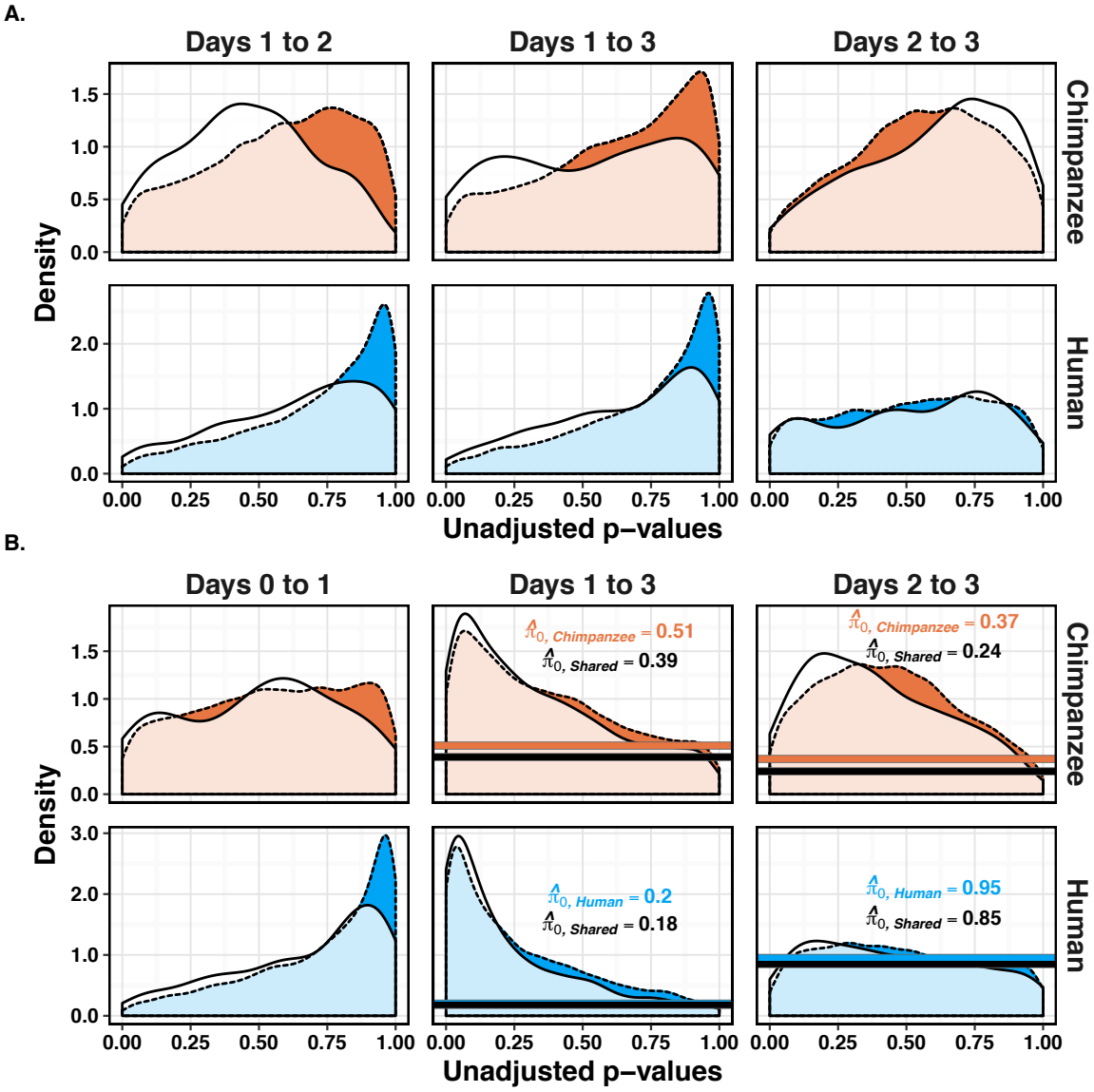


Figure 3.20: **The reduction of variation in gene expression is localized to the primitive streak.** A. *P*value distributions from F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course. Each plot shows the *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).

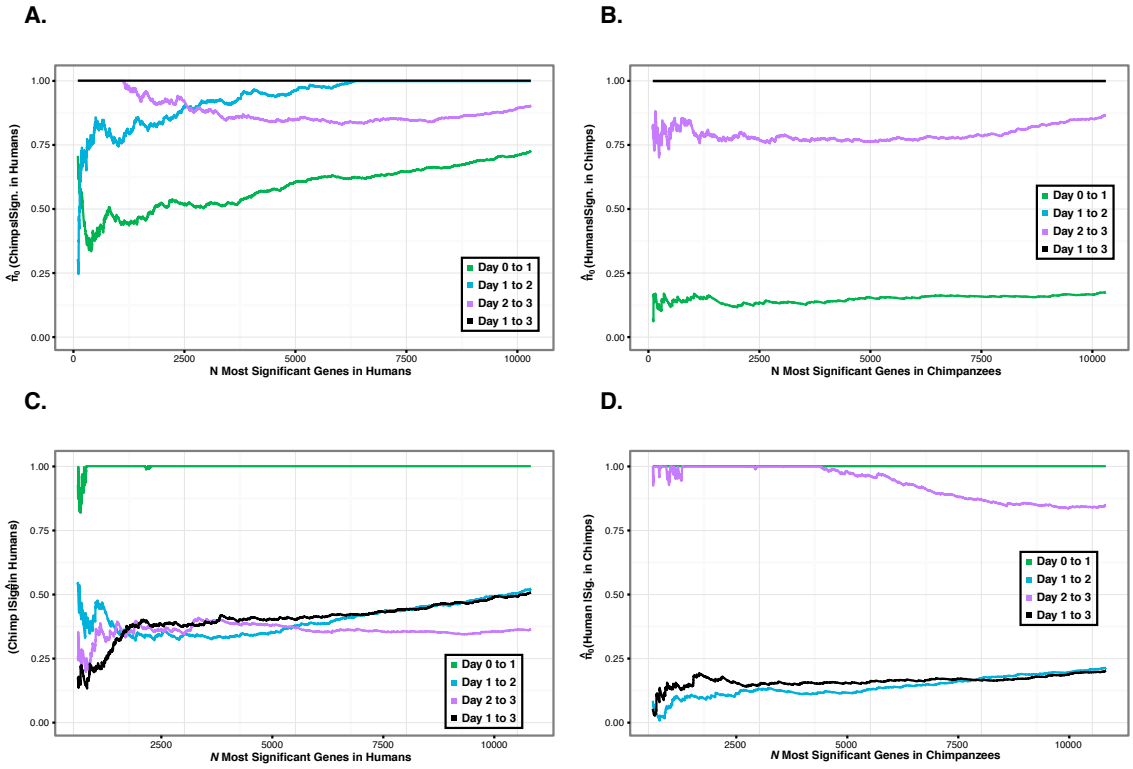


Figure 3.21: **The patterns of change of variation in gene expression are robust with respect to a cutoff based on the number of genes.** *P*value distributions of F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the 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. pi zero 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.

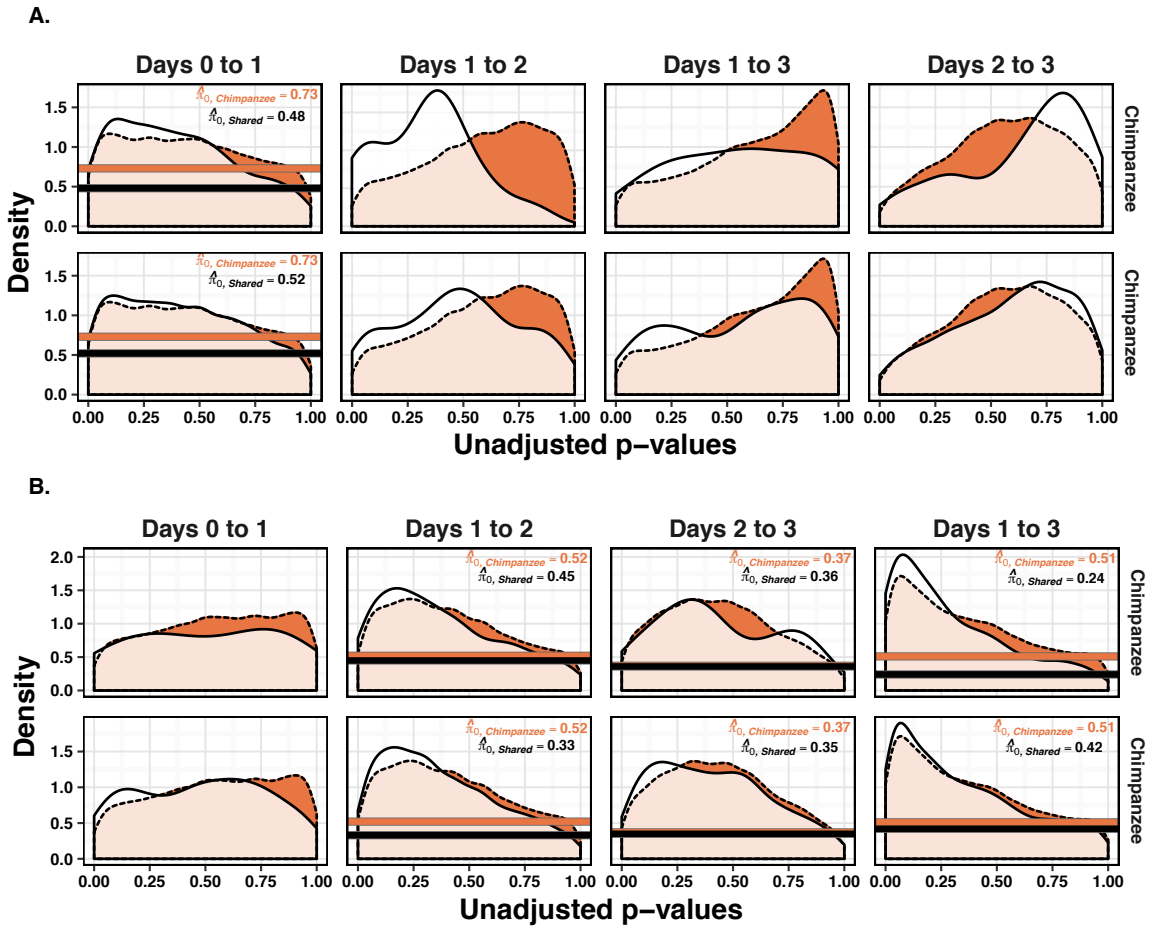


Figure 3.22: **The patterns of change of variation in gene expression are robust with respect to *P*value cutoff in the chimpanzee samples.** Pi zero is the estimated proportion of null tests in each distribution. A. *P*value distributions from F tests against the null hypothesis that there was no reduction 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). Pi zero 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).

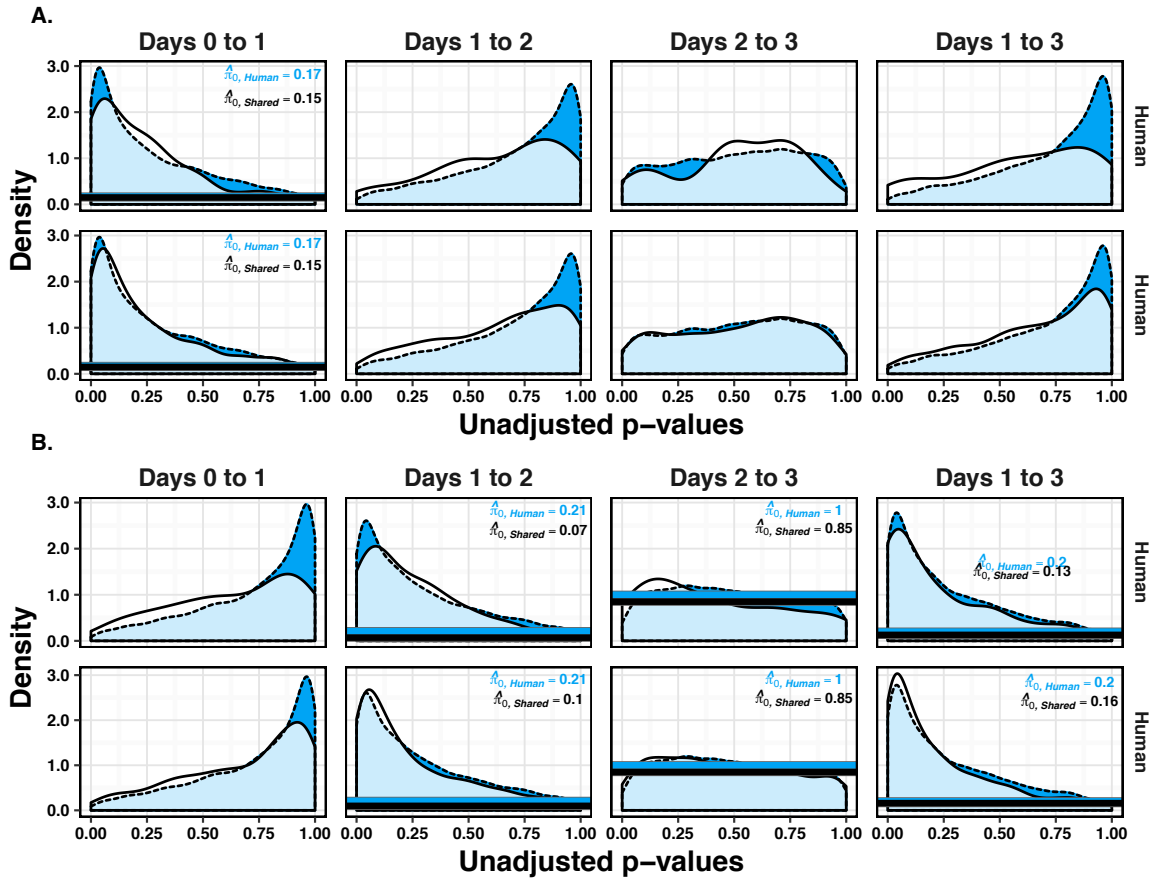


Figure 3.23: **The patterns of change of variation in gene expression are robust with respect to *P*value cutoff in the human samples.** Pi zero is the estimated proportion of null tests in each distribution. A. We plotted the *P*value distributions of F tests against the null hypothesis that there was no reduction in variation in gene expression levels as the samples progress along the time course 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 *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 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).

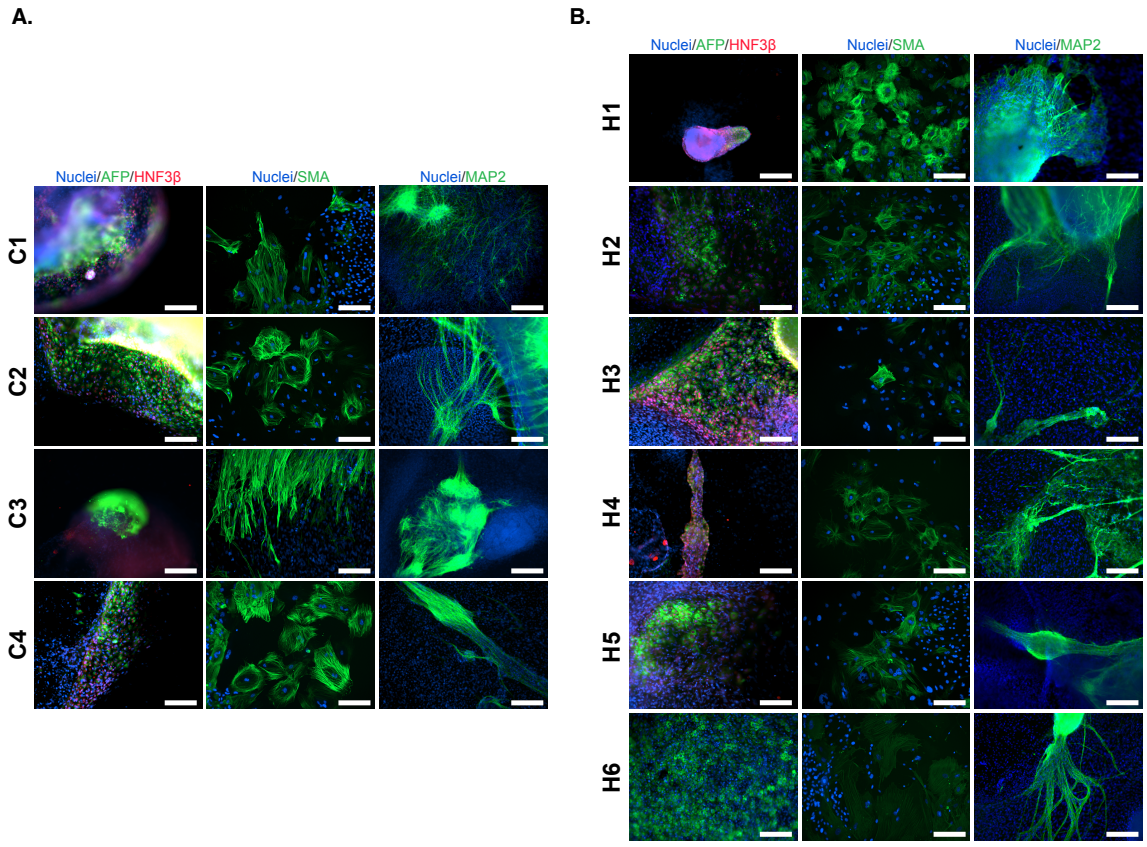


Figure 3.24: **Assay of pluripotency for iPSC lines used in this study.** A. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for chimpanzee iPSC lines. B. Immunocytochemistry (ICC) staining of spontaneously differentiated embryoid bodies for human iPSC lines, antibodies identifying celltypes derived from the three germ layers as indicated.Scale bar: 200 uM.

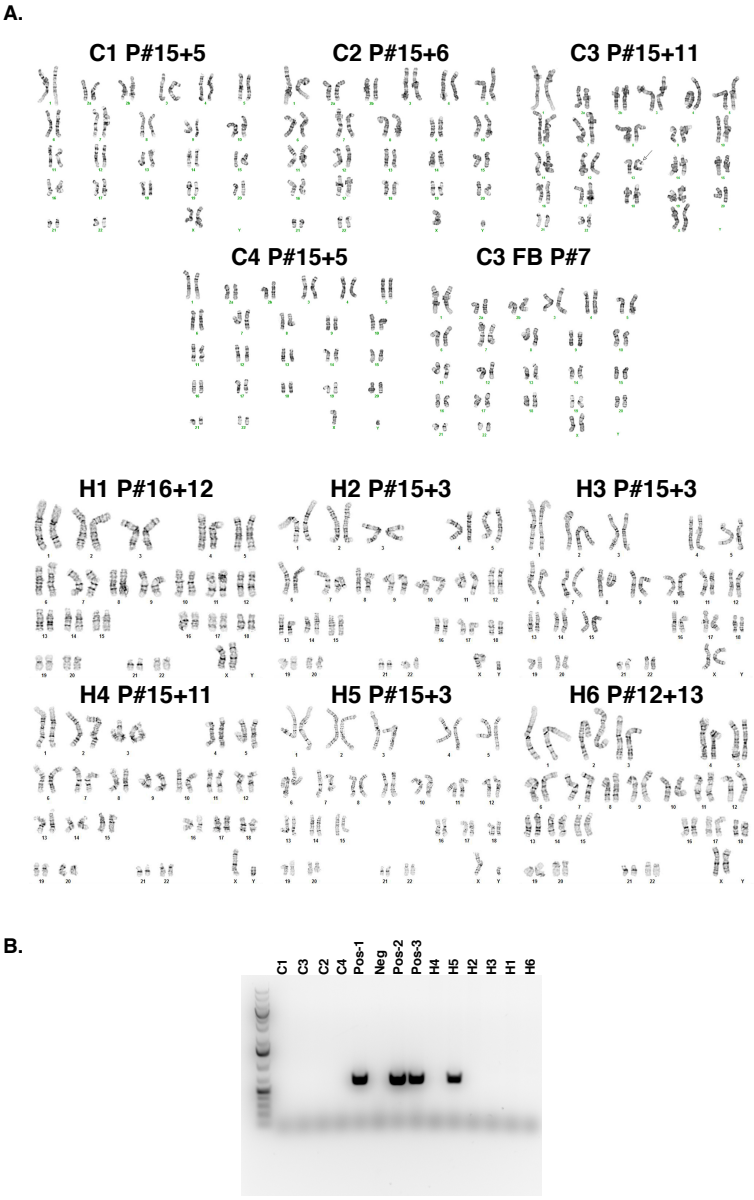


Figure 3.25: **Quality control for the iPSC lines used in this Study.** A. Karyotypes for human and chimpanzee iPSC lines. We identified additional bands in the p-arms of one chromosome 13 homolog and one chromosome 18 homolog for chimpanzee iPSC line C3. Thus we tested the source fibroblast line (C3 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.

3.9 Supplementary Tables

Table 3.1: **Purity estimates for each of the 63 samples.** Column 1 contains the Sample ID and Day. Column 2 contains the purity estimate. Samples without purity estimates were assigned an "NA" designation.

Table 3.2: **RNA Integrity Number (RIN) measured with a Bioanalyzer (Agilent) for each of the 63 samples.** Column 1 contains the Sample ID and Day. Column 2 contains the RIN score. Samples without RIN scores are annotated with an "NA" designation.

Table 3.3: **TMM- and cyclic loess-normalized log2 counts per million for the 10,304 genes analyzed in this study for each of the 63 samples.** Column 1 provides the Ensembl gene name. Row 1 contains the Sample ID and Day (to the left and right of the hyphen, respectively).

Table 3.4: **Biological and technical factors of interest, and the accompanying values for each of the 63 samples.** (A) 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. Rows 4-66 provide the values for each variable for the 63 samples. Row 67 contains information for the outlier sample, H1B. (B) A glossary of the technical factors and their definitions for this study.

Table 3.5: **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. The "pval" refers to the Bonferroni-Hochberg adjusted p-value.

Table 3.6: **Gene Ontology results.** GO when comparing the two most divergent expression patterns, motifs 4 and 7 from Figure 3.5, to the 10,304 genes studied. The output is from the Panther database testing for an overrepresentation of genes in given GO biological process categories, using the Bonferroni correction for multiple testing.

Table 3.7: **The number of interspecies DE genes 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 3.8: **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 (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 3.9: **Distributions of P values from all F tests.** (A) P-value distributions from F tests of reduced variation from day 0 to 1, 1 to 2, 2 to 3, 1 to 3 in each species. (B) P-value distributions from F tests of increased variation from day 0 to 1, 1 to 2, 2 to 3, 1 to 3 in each species.

Table 3.10: **100,000 pi zero values calculated from permuted P-values obtained from F tests for reduced variation in gene expression levels.** Column 1 contains 0 estimates for the chimpanzee samples conditioned on significance in human samples based on permutations (mean = 0.274, standard deviation = 0.046). Column 2 is the same data type but was calculated using human samples conditioned on significance in chimpanzees (mean = 0.825, standard deviation = 0.050).

Table 3.11: **Log-likelihoods for Cormotif models using 7 and 8 correlation motifs.** Columns 1 and 2, respectively; seeds 1-100 were used to generate both lists.

CHAPTER 4

INTEGRATING CLINICAL AND BIOLOGICAL DATA TO PREDICT TREATMENT RESPONSE IN INDIVIDUALS WITH ANOREXIA NERVOSA.

4.1 Abstract[1]

Maintenance and recovery after hospitalization for anorexia nervosa (AN) is challenging, and up to 45% of patients are rehospitalized. Better predictions of outcomes post-discharge–specifically for those at risk for poor outcomes, which are less studied than who does well–could enable earlier and more personalized interventions. To develop predictive models, we used clinical, demographic, and biological information that was available in 29 females with AN at hospital admission, discharge, and 3 timepoints post-discharge from the UNC Eating Disorder Unit. We could not robustly predict body mass index (BMI), change in BMI, or rehospitalization at any time point. Based on these analyses, we suggest a new framework for predicting treatment response in individuals with anorexia nervosa.

Our findings concerning clinical outcome also point to new areas for research. Consistent with previous findings, BMI at discharge is well correlated with BMI post-discharge. Unexpectedly, we found that BMI at discharge showed minimal correlations with rate of BMI change, and this effect was replicated in an independent cohort. These results suggest that change in BMI may play an important role in rehospitalization, in a way that is distinct from absolute BMI.

1. Citation for chapter: Lauren E Blake, Rachel Guerra, Christopher Hubel, Laura M. Thorton, Tae Kim, Yuxin Zou, Patrick F. Sullivan, Cynthia M. Bulik, Jessica H. Baker. Integrating clinical and biological data to predict treatment response in individuals with anorexia nervosa. *Manuscript in prep.*

4.2 Introduction

4.2.1 The importance of predicting outcome for individuals with anorexia nervosa (AN)

Anorexia nervosa (AN) is a heritable eating disorder, in which symptoms include low body weight and an extreme fear of gaining weight [5, 6, 72, 97]. Individuals with AN at extremely low body weights (e.g. ¡ 75% of ideal body weight) are often hospitalized in specialized units for nutritional rehabilitation [5, 85]. After refeeding and weight gain in the hospital, these individuals are discharged to lower levels of care [5, 41]. Maintenance and recovery after hospitalization for AN is challenging, and up to 45% of patients are rehospitalized [43, 186]. Rehospitalization generally occurs within 18 months of discharge [186]. Underscoring the severity and chronicity of disorder, over 20% exhibit a severe and enduring course of AN [67, 104]. Better predictions of outcome post-hospital discharge could enable earlier and more personalized interventions [74]. In particular, understanding those with poor outcomes, who are less studied than individuals that do well, is important [67, 78, 79]. In particular, prediction is a critical step in determining earlier interventions and developing more personalized therapies for AN.

4.2.2 Challenges for predicting outcomes in AN

While prediction has important clinical implications, building robust predictive models of AN outcome has been extremely difficult. One challenge is that the field lacks consensus on which outcome(s) to predict. For example, AN symptoms or diagnosis criteria are frequently used [67, 73]. Post-discharge BMI has also been studied [107], but the clinical utility of BMI in AN is somewhat uncertain [129]. Other studies have attempted to predict "weight relapse", but since there is no consensus definition for weight relapse [114], this endpoint is difficult to study in a reproducible way. Few studies have focused on rehospitalization as

an endpoint [43]. Additionally, behavioral, social, and psychological factors vary in their ability to predict AN outcomes [126]. The lack of subjective predictors has led to the investigation of objective predictors of outcome in individuals with AN [19, 136]. Most studies incorporating biological information have focused on weight or hormone levels during hospitalization [19, 30]. Hormone levels– including leptin, ghrelin, and insulin-like growth factor-1– are not consistently associated with outcome [98, 150, 190, 193, 207]. Whether studies use subjective and objective predictors, most have very small sample sizes (generally <50 individuals) and do not attempt to replicate in another dataset [43, 95, 186]. Both of these factors likely contribute to the lack of reproducibility of these predictors [38].

4.2.3 Gene expression levels as predictors of outcome

Other psychiatric subfields have explored genome-wide measures, including gene expression levels, as objective predictors for disorder diagnosis and treatment [31, 44, 99, 106]. Gene expression levels are tested as biomarkers because they are quantitative, relatively unbiased, and relatively easy to obtain [139]. For an AN cohort, blood-based measurements are particularly attractive, as blood is drawn at admission, at discharge, and during hospitalization as part of the standard of care. Gene expression levels assessed from blood in individuals with AN at admission and discharge was tested in a pilot study [117]. We investigated gene expression levels in whole blood as candidate biomarkers for outcomes within 1 year post-discharge. To develop predictive models, we obtained clinical, demographic, and biological information from 29 females with AN at admission, discharge, and 3 timepoints post-discharge from the Eating Disorder Unit (EDU) at the University of North Carolina. Unfortunately, we could not predict BMI, change in BMI, or rehospitalization at any time point in a replicable manner. However, we do demonstrate a low correlation between discharge BMI and change in BMI post-discharge, which we replicated in an independent cohort. This result has strong implications for the definition of outcome in predictive models. Overall, the results of this

pilot investigation will inform the design of future studies predicting AN patient outcomes.

4.3 Results

4.3.1 Study design and sample characteristics

We performed a longitudinal study of individuals with AN (females over 15 years of age) hospitalized in the EDU (Figure 1A; Methods). We collected information at 5 time points: admission, discharge, 3 months post-discharge, 7 months post-discharge, and 12 months post-discharge (Figure 1B). At EDU admission, we collected extensive clinical and demographic information, as well as whole blood samples from 55 individuals (Figure 1B; Methods). Within 24 hours of collection, whole blood samples were sent for laboratory analysis, including complete blood count (CBC), metabolic panel, and RNA-sequencing (Methods). Each individual underwent inpatient treatment at the EDU for at least 7 days (Figure 1B) [117]. At discharge, we again collected clinical information and a whole blood sample (Figure 1A). At 3, 7, and 12 months post-discharge, we collected clinical information, including weight, eating disorder severity, and treatment status (Figure 1A, 1C-D). Additional collection and patient information can be found in the Methods.

4.3.2 Nutritional rehabilitation treatment induces significant biological changes

We sought to characterize the variation in clinical and gene expression measurements across our patients and timepoints. The majority of patients gained weight during hospitalization, and weight gain increased with time spent in the EDU (Figure 2A). This change was also reflected in the global survey of gene expression variation (Figure 2B; Methods). When we performed a principal component analysis (PCA), we found that the first principal component (PC) was correlated with RIN score (Figure 2B; regression of PC1 by RIN, Pearson's

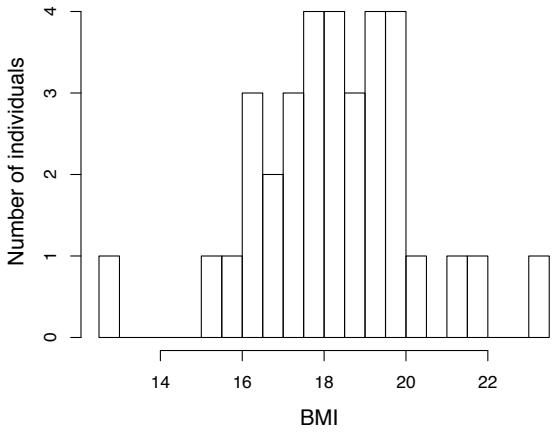
A)

	Admission	Discharge	3 months post-discharge	7 months post-discharge	12 months post-discharge
Number of individuals	55	55	41	38	32

B)

Variable	Mean/SD (Range) or Frequency
Age	25.2 / 8.3 (15-50)
Age of AN onset	16.4 / 5.1 (9-41)
Number of previous hospitalizations	1.5 / 2.8 (0-13)
AN subtype	Restricting = 64%, Binge/purge = 34%, EDNOS = 2%
Weight (lbs)	90.7 / 13.1 (54-121)
BMI at admission	15.1 / 1.8 (10-18)
Days on EDU	30.5/17.2 (7-78)
Discharged against medical advice (AMA)?	No = 80%, Yes = 20%

C)



D)

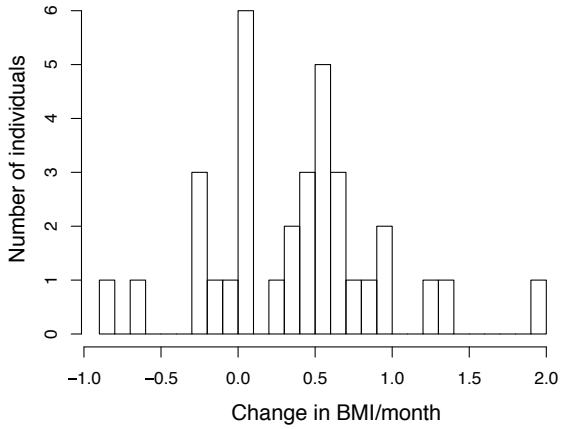


Figure 4.1: **Study design and information about nutritional rehabilitation.** (A) Study design and outcomes measured. (B) Selected sample information taken during hospitalization. (C) BMI at 3-months post discharge (D) Change in BMI/month from discharge over the next 3 months.

r = 0.34). PC2 is correlated with time, as well as age of admission and 5 of the CBC measures (regression of PC2 by time, r = 0.23 and r = 0.24-0.37, respectively). This indicates that sample quality, including cell type heterogeneity, is a major factor for gene expression

variation in our data.

We used a linear model framework to identify gene expression changes between admission and discharge (Methods). We classified 551 genes as differentially expressed (DE) between timepoints (at FDR of 5%; Table 4.1). Using Gene Ontology (GO), we found that many of these genes are enriched in metabolic pathways. When we included weight change as a covariate, only 8% of these genes remain DE, suggesting a strong genomic response to nutritional rehabilitation and, potentially, subsequent weight gain.

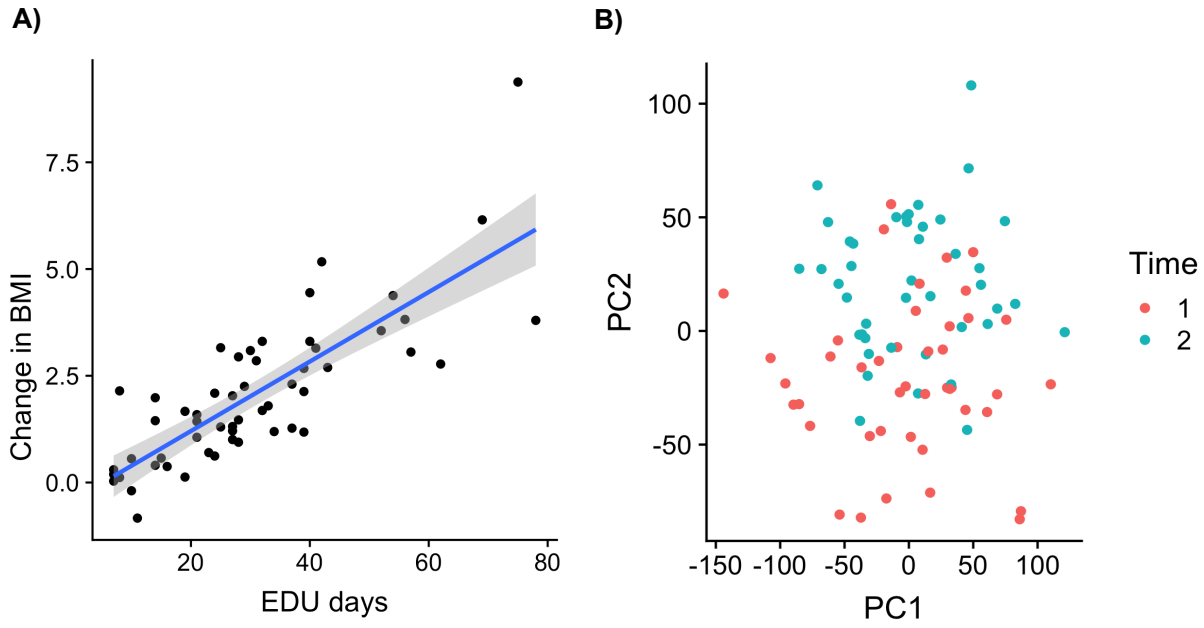


Figure 4.2: **Biological samples at admission and discharge.** (A) Patient weight gain is highly correlated with number of days in the EDU. The blue line is a best-fit linear regression line. Grey shading is the 95% confidence interval. (B) A global survey of gene expression variation (n = 43 individuals in the 3 months post-discharge).

4.3.3 Characterizing post-discharge outcomes

We then sought to understand the impact of clinical variation on patient outcome. We considered three metrics for post-discharge outcome: BMI, change in BMI/month, and re-hospitalization (Figure 1C-D, Methods). Most individuals in our study continued to increase

in BMI post-discharge, particularly in the first 7 months post-discharge (84%, 79%, and 75%, at 3-, 7-, and 12-months post-discharge, respectively). The average change in BMI/month were also positive. Unsurprisingly, a larger percentage of individuals were rehospitalized due to eating disorder symptoms later in the time-course than the first 3 months (15-31%).

We next focused on outcomes forindividuals that discharged against medical advice (AMA), a group of particular interest to AN clinicians. Interestingly, individuals gained similar amounts of weight in the hospital regardless of AMA discharge status. However, individuals that discharged AMA were more likely to lose weight in the next 3 months than individuals that did not discharge AMA ($P < 0.05$, Figure 3).

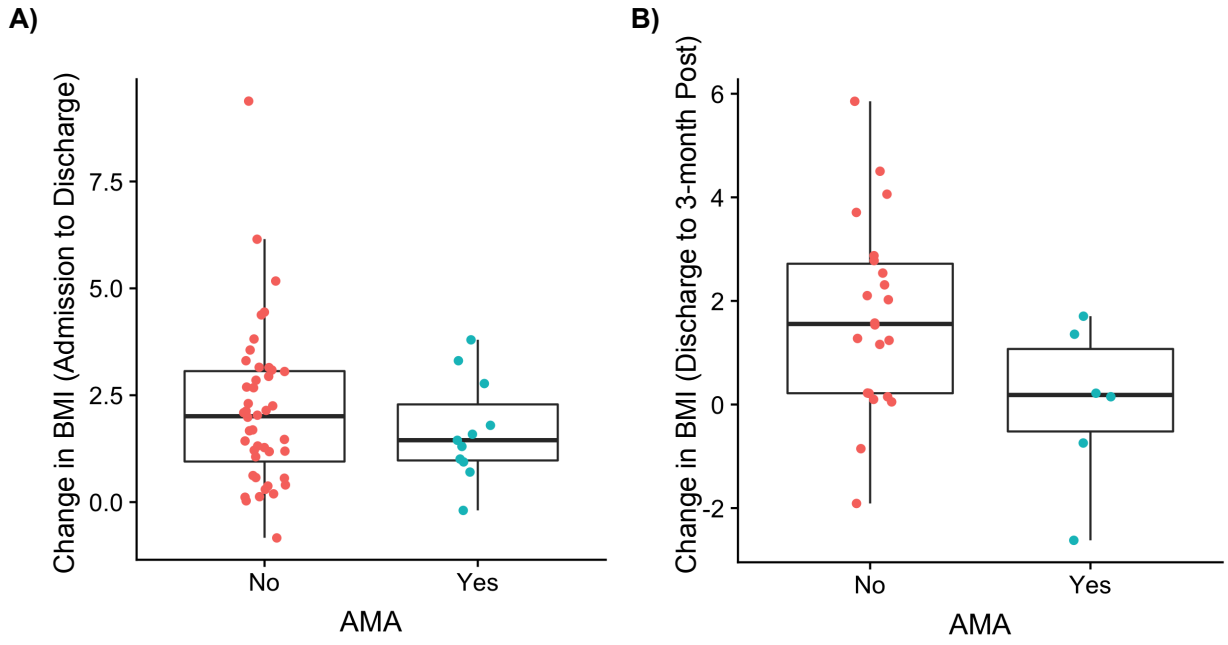


Figure 4.3: **AMA Discharge Status** Individuals that discharged AMA were more likely to lose weight over the next 3 months than individuals that did not discharge AMA..

4.3.4 Variation in gene expression levels does not robustly predict variation in any outcome

Finally, we sought to predict patient outcome based on clinical and gene expression data. We implemented a support vector machine-based model from machine learning to predict individual outcomes using gene expression profiles [2].Our model contained 1000 genes that the expression levels were most highly associated with, along with 8 additional clinical and technical variables (Methods). Due to extreme physiological changes during nutritional rehabilitation (Figure 2A), we included only discharge gene expression levels. We would ideally have accounted for overfitting by building the predictive model on our data set and then validated this model on a data set collected independently. However, we were unable to acquire an analogous data set with gene expression levels and outcomes post-discharge. Therefore, we initially focused onBMI 3 months post-discharge. Specifically, we constructed a model using 29 individuals for which we had data from each time point. We then tested the model on an additional 12 individuals from our study with missing data at later time points (Figure 1A). Using this strategy, we were unable to reliably predict BMI 3 months post discharge ($R < 0$). We repeated this process using change in BMI/month over the first 3 months post-discharge, and also could not predict this outcome. We attribute these results to a small sample size.

In a model without gene expression levels, BMI at discharge was a significant predictor for BMI 3 months post-discharge (Figure 4A), but not change in BMI/month over the same time period (Figure 4B). When we investigated this trend further, we found that BMI at discharge is lowly correlated with rate of BMI change ($r = 0.19$, 0.10, -0.05, respectively, $P > 0.29$ for each time point). This pattern could not be explained by rehospitalization status (Figures 4B-C). This low correlation replicated in an independent cohort of individuals with AN treated at the University of Pittsburgh Medical Center (Methods, $P > 0.35$). Overall, these results suggest that discharge BMI alone could be used to predict BMI range post-

discharge, but can not be used to determine whether an individual will increase or decrease in BMI post-discharge.

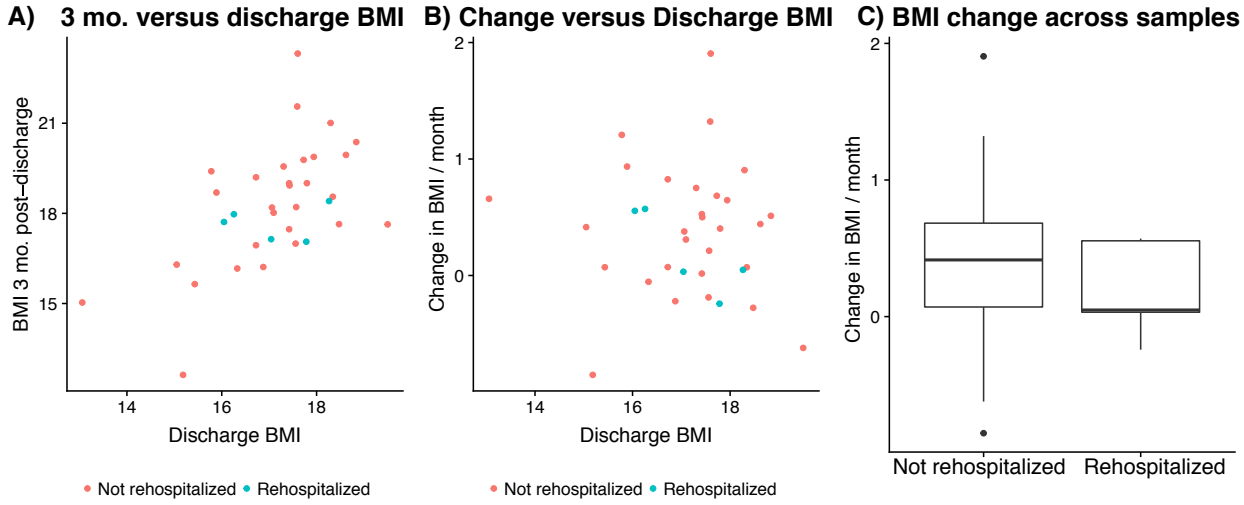


Figure 4.4: **Relationship between BMI at discharge and outcomes.** (A) Consistent with previous findings, BMI at discharge is well correlated with BMI post-discharge (Pearson's r = 0.60, 0.53, 0.41 for 3-, 7- and 12-months post-discharge, respectively, P <0.01 for each). (B) BMI at discharge is lowly correlated with BMI change/month (r = 0.19, 0.10, -0.05, respectively, P >0.29 for 3-, 7-, and 12-months post-discharge), which replicated in an independent cohort from the University of Pittsburgh. (C) As expected, the increase in BMI/month is greater in individuals that were not rehospitalized within 3 months than those people that were rehospitalized.

4.4 Discussion

Over 20% of individuals hospitalized with AN are re-hospitalized within 1 year [43]. This cycle creates a "revolving door" phenomenon in care that is financially and psychologically burdensome for patients and families. Many studies have identified differences in clinical and biological factors that are associated with differences in outcomes [79, 94, 119]. Therefore, a natural next step is predicting post-discharge outcome. Indeed, robust prediction is extremely important for earlier and more targeted intervention. In addition, predictive models

could help determine the level of care needed after hospital discharge and be used to educate patients and families about their personal risk for a given outcome.

Despite its potential to impact patient care, developing robust predictive models of AN outcome has proven exceptionally difficult. This lack of robustness could be driven by challenges with outcome (response variables), predictor variables, or study design. Therefore, our original goal was to identify genes whose expression levels were predictive of outcome within 1 year of hospital discharge. While we failed to meet this objective, this pilot study highlights key considerations for future AN outcome prediction studies.

4.4.1 Prioritize defining measurable, meaningful, and clinically actionable outcomes

A response variable should be measurable, meaningful and clinically actionable. For example, in the field of cardiology, there is a calculator for the 10-year risk of heart attack [86]. In AN, AN diagnosis, weight, or BMI post-discharge are used [79]. Although BMI can be easily measured, the meaning is less clear [147].For BMI post-discharge to be used as an outcome, first a reproducible relationship between BMI and AN severity should be established.

Recognizing that change in BMI might be more important than absolute BMI, other studies have attempted to predict "weight relapse". However, there is no consistent definition of weight relapse in the field [114]. Rather than pick a definition of weight relapse, we decided to predict change in BMI/month. Admittedly, this value can be easily calculated but its implications are currently unknown.

There is some literature to suggest change in weight is meaningful. For example, preliminary work suggests that a rapid increase in weight post-discharge is more highly correlated with BMI after 1 year [132]. Consistent with previous studies [28, 138], BMI at discharge is significantly associated with BMI post-discharge but not change in BMI/month (Figure 4). These results suggest that we can currently predict the approximate BMI of an individual

post-discharge, but not whether patients will increase or decrease weight over time. This is a subtle but important distinction because it means that factors that predict post-discharge BMI (such as discharge BMI) may not predict the direction of change in BMI. Future studies that set out to predict BMI post-discharge should also attempt to predict change in BMI post-discharge.

An outcome that may be clinically actionable is risk of rehospitalization [43, 95, 186]. Rehospitalization is a clearly-defined endpoint. Furthermore, earlier and more frequent intervention at lower levels of care could help prevent rehospitalization. While higher risk of rehospitalization may be clinically actionable, more research needs to be done to assess the reproducibility of rehospitalization risk. This work is particularly important in the United States, due to the influence of health insurance plans on whether an individual with AN enters inpatient treatment. To decrease the influence of insurance on outcome, one area to explore could be the risk of a particular eating disorder related symptoms, such as body mineral density or the risk for a cardiac incident [74, 105, 141, 175].

4.4.2 The importance of study design, including replication

Even if we are able to find an appropriate outcome, current sizes of AN studies make it difficult to establish robust relationships between predictors and outcomes. Therefore, it is important for eating disorder researchers to collect the same predictors (phenotypes) and outcome information. This consistency will allow for replication using data from multiple centers, as we did with data from the University of Pittsburgh Medical Center. It is also important for metadata (including technical variables) to be collected and made available, particularly in eating disorder genomic studies [103].

Another course of action to be implemented in parallel is to use larger scale data, such as those from electronic health records or national registries, to generate hypotheses [173]. Then, these hypotheses could be rigorously tested in smaller, more well controlled studies

with deep phenotype information.

Overall, without progress in the areas of outcomes and study design, it is unclear whether any collected biological measure (e.g. hormone levels, gene expression levels, etc.) will robustly predict AN outcome post-discharge. Working through these challenges will enable prediction that impacts the care of individuals with AN.

4.5 Methods

4.5.1 Participants and Ethics Approval

Participants were females over 15 years old who met DSM-5 Criteria for AN. All individuals were admitted for inpatient treatment at the Eating Disorder Unit (EDU) at the University of North Carolina at Chapel Hill Center of Excellence for Eating Disorders. This study was approved by the Biomedical Institutional Review Board at the University of North Carolina at Chapel Hill. All participants provided written informed consent. For information regarding the replication data collected at the University of Pittsburgh Medical Center, see [28].

4.5.2 Clinical and Whole Blood Information

Individual's weight and height was collected during hospitalization, and BMI calculated, in the same manner as [117]. Additional information was collected from patient surveys and electronic health record (EHR) access. Post-discharge follow up information was collected via phone calls by trained research assistants. Blood draws were made as described in [117]. 1 of the 2 tubes of whole blood was mailed to the NIMH Center for Collaborative Genomics on Mental Health Disorders at Rutgers University for RNA extraction and processing. The University of North Carolina High Throughput Sequencing Facility added barcode adaptors

and sequenced the 50 base pair paired-end RNA-seq libraries on the Illumina HiSeq 2500 on two flowcells. After sequencing, we confirmed that the raw RNA-seq reads of high quality using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

4.5.3 Quantification, normalization, transformation, and QC of the RNA-seq reads

As recommended by the Lineberger Bioinformatics Core, reads were mapped to hg38 (GRCh38.p2, Gencode Release 22) using STAR (v2.6.0a) [60, 61] and quantified these transcripts from these files using Salmon (version 0.8) [154]. The R/Bioconductor package tximport was used to associate transcripts to genes [184]. TMM-normalized log2(CPM) gene expression values were calculated [121]. To assess the data quality, we performed Principal Components Analysis (PCA) on the normalized log2(CPM) values. After recording clinical, biological, demographic, and technical factors, we explored the relationship between these factors and the PCs in the same manner as [21].

4.5.4 Pairwise differential expression (DE) analysis and weighted gene co-expression network (WGCNA) construction

Differential expression analysis was performed using a linear model-based empirical Bayes method using the R packages limma+voom [121, 164, 180]. For the initial pairwise differential expression comparison, the timepoint and 8 other variables were modeled as fixed effects, and individual as a random effect. A second comparison was made with the same model, but also including weight change as a fixed effect. For each pairwise DE test, multiple testing was corrected for using the Benjamini and Hochberg false discovery rate [18]. Gene Ontology (GO)[4] was implemented as described in [21]. The CRAN/R package WGCNA [120] was utilized to construct networks for each set of DE genes (FDR-adjusted P value <0.01), using

the default parameters.

4.5.5 Choice of outcomes

First, we needed to determine which outcome to predict. We initially considered looking at trajectories. These trajectories were inconsistent (e.g. very few individuals that consistently gained consistently or lost during the 1 year post-discharge). Then, we considered moments of the BMI distribution from discharge to 12 months post-discharge. Average weight (discharge to 12 months post-discharge) is hard to interpret clinically. Due to weight fluctuations and lack of clinical support, we chose not to construct a cut off for the BMI change that would be considered significant. Therefore, we chose to focus on BMI at 3-, 7-, and 12-months post-discharge and change in BMI/month between these three time points as target variables.

4.5.6 Attempts to build predictive models

For each outcome, we used LASSO regression to determine the top 1000 genes in which variation in expression levels were most predictive of the particular outcome. We also included 8 additional clinical and technical variables in our support vector machine-based models. The clinical and technical covariates were significantly associated ($P < 0.05$) with discharge gene expression levels PCs 1, 2, 3, 4, or 5: age at admission, absolute monophil count, absolute neutrophil count, absolute eosinophil, absolute leukocyte count, RIN score, and binge eating status. We used the 29 individuals for which we had data from each time point in the models. The support vector machine (SVM)-based models were generated using the R package e1071 (https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf) [101, 2]. The models were tested on the additional 12 individuals from the 3 months post-discharge timepoint.

4.6 Acknowledgements

I thank members of the Gilad and Stephens labs and the UNC Center of Excellence for Eating Disorders for helpful discussions and comments on the manuscript. I also thank the participants and their families for participation in this study.

4.6.1 Author contributions

JHB conceived of the study and designed the experiments, with input from PFS and CMB. RG and JHB collected the data. LEB analyzed the data, with input from CH, LMT, PFS, TK, YZ, and CMB. LEB and JHB wrote the paper.

4.7 Supplementary Tables

Table 4.1: **Information about genes DE between admission and discharge.**

CHAPTER 5

CONCLUSION

5.0.1 Eating disorder genetics

Notably, the goal of human genetics is to better understand human complex traits and diseases. As described in previous chapters, a properly designed study can facilitate this biological understanding. In Chapter 2, the use of matched tissues between species allowed us to estimate the contribution of DNA methylation to conserved inter-tissue differences in gene expression. We also found that conserved gene regulatory differences are more likely to be involved in gene regulatory networks and protein-protein networks than non-conserved differences. In Chapter 3, we leveraged matched induced Pluripotent Stem Cell (iPSC) panels from humans and chimpanzees. In this developmental timecourse from iPSCs to the endoderm stage, my collaborators and I found high conservation during the differentiation. Furthermore, we reported a reduction in regulatory variation at the primitive streak stage in a significant number of genes. Due to our study design and careful analysis of technical variables, we were able to provide evidence that this canalization is driven by biological, rather than technical, factors. Finally, in Chapter 4, I discussed power to identify biologically-based predictors in the context of a longitudinal study design.

Indeed, study design plays a critical role in all types of comparative studies, whether the goal is to understand differences between tissues, species, or diseased and healthy individuals. In particular, genomic studies can be very useful for understanding complex traits and diseases, including eating disorders. When designed properly, genomic studies can uncover biological mechanisms and may be used for prediction related to the disorder and treatment. For example, understanding the underlying biological mechanisms for a given disorder can help to develop new drug targets [195]. This course of action is particularly important for eating disorders, as few medications are available specifically for BN or BED, and no phar-

macological interventions have been approved specifically for AN treatment [174]. This line of research could also help the field gain a greater understanding of how existing pharmacological interventions work [195]. There may be treatments that could be developed to ease the process of recovery for AN patients, including nutrient absorption and the pain experienced during nutritional rehabilitation [117].

Knowing that the treatments are important, anoter developing area within eating disorders is that of prediction. Determining risk of eating disorders, responses to treatment, and risk for relapse has major implications for screening [114]. Integrating genomics into one's risk profile could increase prediction performance metrics, which will ultimately increase patient care. Prediction is also important for personalized medicine. Currently, individuals with eating disorders are often treated with therapeutics for comorbid conditions. Unfortunately, these drugs are not tested in very low weight individuals, such as those with AN. Since weight influences metabolism, some clinicians use genetics to inform which therapeutic to prescribe [36]. As additional therapeutics become available for the treatment of eating disorders and common comorbidities, individuals could be prescribed an intervention(s) shown to be most effective for their clinical and genetic profile.

Our efforts to understand the biology of eating disorders will have a profound social and psychological impact on eating disorder patients and their families. Like many psychiatric disorders, eating disorders are both highly stigmatized and poorly understood [36]. Therefore, diagnoses often lead to feelings of shame and guilt [65]. These feelings can be alleviated by understanding the neurobiological and genetic causes for these disorders, providing considerable and immediate relief to patients and their families [36].

Legitimate challenges exist for the field. Therefore, I will now describe recent progress in eating disorder genomics and consider key challenges in uncovering the biological mechanisms of eating disorders and in predicting eating disorder outcomes. From my discussion of these challenges, I also identify opportunities to improve future genomic studies of eating disorders.

5.0.2 Progress on the use of genetics to understand the biological mechanisms of eating disorders

At this point, research into the genetic risk factors of AN and other eating disorders is still in its early stages [36]. Furthermore, the majority of eating disorder research has focused on AN. The second and most recent genome wide association study (GWAS) of AN found the heritability of AN to be 11-17% [215]. In this study of over 16,000 AN cases and 55,000 eating disorder controls, 8 SNPs were associated with differences between AN cases and controls [215]. Some of these SNPs associated with AN cases have also been associated with metabolic traits, such as body mass index (BMI) [215].

5.0.3 Challenges and possible solutions going forward

I believe that the AN GWAS represents major progress. It, however, also highlights many of the challenges to the field going forward. All 8 significant hits were located in non-protein coding regions [216]. Consequently, it is hard to dissect the impact of these SNPs on disease risk. This interpretation is particularly difficult because these genes may impact multiple tissues, or even have different effects in different tissues. Unfortunately, the functional follow-up for AN is incredibly difficult, as it is difficult to access human brain tissues and there are no widely accepted mouse models of AN [174].

To overcome challenges around access to human brain tissues and functional followup, researchers in other psychiatric fields have generated iPSC-based models from cases and controls [33, 34]. A preliminary study demonstrated that it is possible to derive neurons from iPSCs in AN cases and controls [149]. However, interpretation again is difficult: it hard to know what phenotype to measure in these "neurons in a dish", particularly for AN. While this original study focused on gene expression levels, iPSC-derived neurons from individuals with schizophrenia have been assessed for other properties as well [32].

Historically, the eating disorder field, and AN in particular, has also faced challenges

around comorbities. For example, there are well documented phenotypic comorbidities between eating disorders and other psychiatric disorders [145]. These phenotypic correlations are also reflected by strong genetic correlations [215]. Furthermore, AN is confounded with low BMI [6]. Since BMI and disease state are confounded, it is difficult to direct assess claims about the metabolic etiology of AN. To attempt to overcome the strong relationship between BMI and AN, [215] compares AN patients to approximately 1,500 constituently thin individuals. While this approach is a good first step, I argue that this approach will be limited in the long term, particularly because there is little transcriptomic, proteomic, and metabolomic data on individuals with AN [117]. To make significant gains on this front, more work needs to be done to understand the multi-faceted contributors to weight more generally.

Although comorbidities complicate the study of eating disorders, the application of newer analytical methods leverage the complexities. For example, to partition variants associated with a psychiatric disorder specifically, compared to more generally, case-case GWAS are currently applied to other psychiatric disorders [45]. Case-case GWAS may be useful to compare across eating disorders. However, a more interesting application could be to apply this technique to individuals with a given sub-phenotype, such as individuals with anxiety but no history of eating disorders compared to individuals with anxiety and eating disorders. The results of this type of study may be helpful for differentiating the multiple components of eating disorders. Overall, understanding the underlying mechanisms of disorders as complex as eating disorders will require substantial long-term investment and multiple channels of inquiry.

5.0.4 Progress on the integration of genetics in predictive models in individuals with eating disorders

Notably, prediction represents a more immediately clinically actionable path. While many studies have attempted to predict long term outcome for individuals with eating disorders, to my knowledge, none have integrated non-hormone-based genetic information [51, 53, 75, 114, 119]. Instead, genetic information has been used to predict risk of AN, rather than long term outcome, primarily through polygenic risk score (PRS) [195]. In the most recent AN GWAS, PRS was applied to explain 1.7% of phenotypic variation [215]. This low percentage of variance explained suggests a current limited clinical utility. PRS has not been calculated for different subtypes or outcomes of AN, which is likely due to sample size. Even if the PRS is higher for these responses, it will be important to integrate clinical and other phenotypic information into the predictors.

5.0.5 Challenges and possible solutions going forward

Knowing the current state of PRS, one of my fears for using this value for disorder risk is the potential for misinterpretation and misuse [170]. Individuals may over interpret an increased risk for a disorder as genetic determinism, which is highly problematic for an already stigmatized class of disorders such as eating disorders [7, 8, 29, 96]. Therefore, I argue that it is more beneficial to focus on individuals that have already been diagnosed with an eating disorder. In these individuals, prediction could directly impact treatment options.

Yet most outcome studies in AN do not take into account genetic factors (including hormones), as described in Chapter 4. These studies face difficult, but not unsurmountable challenges related to study design, defining outcomes, choice of predictors, and analysis methods. As argued in Chapter 4, the first priority should be to determine more rigorous and clinically relevant definition of outcomes for eating disorders, such as quantifying risk of

rehospitalization for individuals with AN.

Overall, the principles learned from more mature fields could be serve as useful guides for the field of eating disorder genomics. For example, extensive record keeping of both biological and technical variables, such as that discussed in Chapter 2, is just beginning to gain traction in the eating disorders genomics field [103]. Furthermore, measures of cellular heterogeneity were extremely important to account for in the iPSC study from Chapter 3. Controlling for cell type heterogeneity is critical when using whole blood from patients, such as in Chapter 4 and [117]. These two examples highlight the importance of cross-talk between multiple genomics fields.

5.0.6 Final words

Ultimately, progress in complex disorders such as eating disorders can only be achieved by integrating "big data" sources, including genomics, into well designed studies. To do so will require a major commitment by federal funding and other major funding agencies to support basic scientists and subsequently, strong collaborations between basic and clinician scientists.

References

[1] 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.

[2] Alexandros Karatzoglou and Alexandros Smola and Kurt Hornik and Achim Zeileis. kernlab - an s4 package for kernel methods in r. *Journal of Statistical Software, Articles*, 11(9):1–20, 2004.

[3] S. J. Arnold, U. K. Hofmann, E. K. Bikoff, and E. J. Robertson. Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and endoderm specification in the mouse. *Development*, 135(3):501–11, 2008.

[4] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, G Sherlock, and Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[5] American Psychiatric Assocation. *Diagnostic and Statistical Manual Of Mental Disorders, fifth edition*.

[6] E. Attia, A. E. Becker, R. Bryant-Waugh, H. W. Hoek, R. E. Kreipe, M. D. Marcus, J. E. Mitchell, R. H. Striegel, B. T. Walsh, G. T. Wilson, B. E. Wolfe, and S. Wonderlich. Feeding and eating disorders in dsm-5. *Am J Psychiatry*, 170(11):1237–9, 2013.

[7] Jehannine C. Austin, Catriona Hippman, and William G. Honer. Descriptive and numeric estimation of risk for psychotic disorders among affected individuals and relatives: Implications for clinical practice. *Psychiatry Research*, 196(1):52–56, 2012.

[8] S. Austin. A perspective on the patterns of loss, lack, disappointment and shame encountered in the treatment of six women with severe and chronic anorexia nervosa. *J Anal Psychol*, 54(1):61–80, 2009.

[9] C. C. Babbitt, O. Fedrigo, A. D. Pfefferle, A. P. Boyle, J. E. Horvath, T. S. Furey, and G. A. Wray. Both noncoding and protein-coding rnas contribute to gene expression evolution in the primate brain. *Genome Biol Evol*, 2:67–79, 2010.

[10] V. A. Balashova and K. M. Abdulkadyrov. [cellular composition of hemopoietic tissue of the liver and spleen in the human fetus]. *Arkh Anat Gistol Embriol*, 86(4):80–3, 1984.

[11] KV Ballman, DE Grill, AL Oberg, and TM Therneau. Faster cyclic loess: normalizing rna arrays via linear models. *Bioinformatics*, 20(16):2778–2786, 2004.

[12] N. E. Banovich, X. Lan, G. McVicker, B. van de Geijn, J. F. Degner, J. D. Blischak, J. Roux, J. K. Pritchard, and Y. Gilad. Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, 10(9):e1004663, 2014.

[13] 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.

[14] S. Barbash and T. P. Sakmar. Brain gene expression signature on primate genomic sequence evolution. *Sci Rep*, 7(1):17329, 2017.

[15] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, and B. J. Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–93, 2012.

[16] D. P. Barlow. Methylation and imprinting: from host defense to gene regulation? *Science*, 260(5106):309–10, 1993.

[17] LB Barreiro, JC Marioni, Blekhman R, Stephens M, and Gilad Y. Functional comparison of innate immune signaling pathways in primates. *Plos Genetics*, 6(10):e1001249. doi: 10.1371/journal.pgen.1001249, 2010.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[19] L. A. Berner, J. A. Shaw, A. A. Witt, and M. R. Lowe. The relation of weight suppression and body mass index to symptomatology and treatment response in anorexia nervosa. *J Abnorm Psychol*, 122(3):694–708, 2013.

[20] J. A. Blake, J. E. Richardson, C. J. Bult, J. A. Kadin, J. T. Eppig, and Group Mouse Genome Database. Mgd: the mouse genome database. *Nucleic Acids Res*, 31(1):193–5, 2003.

[21] L. E. Blake, S. M. Thomas, J. D. Blischak, C. J. Hsiao, C. Chavarria, M. Myrthil, Y. Gilad, and B. J. Pavlovic. A comparative study of endoderm differentiation in humans and chimpanzees. *Genome Biol*, 19(1):162, 2018.

[22] R. K. Blashfield and A. K. Fuller. Predicting the diagnostic and statistical manual of mental disorders (fifth edition): The mystery of how to constrain unchecked growth. *J Nerv Ment Dis*, 204(6):415–20, 2016.

[23] 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.

[24] 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.

[25] Ran Blekhman. A database of orthologous exons in primates for comparative analysis of rna-seq data. *Nature Precedings*, page doi: 10.1038/npre.2012.5360.4, 2012.

[26] 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.

[27] C. Bock. Analysing and interpreting dna methylation data. *Nat Rev Genet*, 13(10):705–19, 2012.

[28] L. P. Bodell, S. E. Racine, and J. E. Wildes. Examining weight suppression as a predictor of eating disorder symptom trajectories in anorexia nervosa. *Int J Eat Disord*, 49(8):753–63, 2016.

[29] K. Borle, E. Morris, A. Inglis, and J. Austin. Risk communication in genetic counseling: Exploring uptake and perception of recurrence numbers, and their impact on patient outcomes. *Clin Genet*, 2018.

[30] H. K. Boyd, L. P. Bodell, K. M. Jennings, A. K. Graham, R. D. Crosby, and J. E. Wildes. Relationship between desired weight constructs and eating disorder severity following treatment for anorexia nervosa. *International Journal of Eating Disorders*, 51(8):870–878, 2018.

[31] J. Breitfeld, C. Scholl, M. Steffens, G. Laje, and J. C. Stingl. Gene expression and proliferation biomarkers for antidepressant treatment resistance. *Transl Psychiatry*, 7(3):e1061, 2017.

[32] K. Brennand, J. N. Savas, Y. Kim, N. Tran, A. Simone, K. Hashimoto-Torii, K. G. Beaumont, H. J. Kim, A. Topol, I. Ladran, M. Abdelrahim, B. Matikainen-Ankney, S. H. Chao, M. Mrksich, P. Rakic, G. Fang, B. Zhang, 3rd Yates, J. R., and F. H. Gage. Phenotypic differences in hipsc npcs derived from patients with schizophrenia. *Mol Psychiatry*, 20(3):361–8, 2015.

[33] K. J. Brennand, A. Simone, J. Jou, C. Gelboin-Burkhart, N. Tran, S. Sangar, Y. Li, Y. Mu, G. Chen, D. Yu, S. McCarthy, J. Sebat, and F. H. Gage. Modelling schizophrenia using human induced pluripotent stem cells. *Nature*, 473(7346):221–5, 2011.

[34] K. J. Brennand, A. Simone, N. Tran, and F. H. Gage. Modeling psychiatric disorders at the cellular and network levels. *Mol Psychiatry*, 17(12):1239–53, 2012.

[35] P. Brodal. Principles of organization of the corticopontocerebellar projection to crus ii in the cat with particular reference to the parietal cortical areas. *Neuroscience*, 10(3):621–38, 1983.

[36] C. M. Bulik, L. Blake, and J. Austin. Genetics of eating disorders: What the clinician needs to know. *Psychiatr Clin North Am*, 42(1):59–73, 2019.

[37] C. M. Bulik, P. F. Sullivan, and K. S. Kendler. An empirical study of the classification of eating disorders. *Am J Psychiatry*, 157(6):886–95, 2000.

[38] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafo. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, 14(5):365–76, 2013.

[39] M. Caceres, J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo, D. H. Geschwind, D. J. Lockhart, T. M. Preuss, and C. Barlow. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A*, 100(22):13030–5, 2003.

[40] 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.

[41] K. Campbell and R. Peebles. Eating disorders in children and adolescents: state of the art review. *Pediatrics*, 134(3):582–92, 2014.

[42] J. A. Capra, G. D. Erwin, G. McKinsey, J. L. Rubenstein, and K. S. Pollard. Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci*, 368(1632):20130025, 2013.

[43] J. Castro, A. Gila, J. Puig, S. Rodriguez, and J. Toro. Predictors of rehospitalization after total weight recovery in adolescents with anorexia nervosa. *Int J Eat Disord*, 36(1):22–30, 2004.

[44] A. Cattaneo, N. Cattane, V. Begni, C. M. Pariante, and M. A. Riva. The human bdnf gene: peripheral gene expression and protein levels as biomarkers for psychiatric disorders. *Transl Psychiatry*, 6(11):e958, 2016.

[45] A. W. Charney, D. M. Ruderfer, E. A. Stahl, J. L. Moran, K. Chambert, R. A. Belliveau, L. Forty, K. Gordon-Smith, A. Di Florio, P. H. Lee, E. J. Bromet, P. F. Buckley, M. A. Escamilla, A. H. Fanous, L. J. Fochtmann, D. S. Lehrer, D. Malaspina, S. R. Marder, C. P. Morley, H. Nicolini, D. O. Perkins, J. J. Rakofsky, M. H. Rapaport, H. Medeiros, J. L. Sobell, E. K. Green, L. Backlund, S. E. Bergen, A. Jureus, M. Schalling, P. Lichtenstein, P. Roussos, J. A. Knowles, I. Jones, L. A. Jones, C. M. Hultman, R. H. Perlis, S. M. Purcell, S. A. McCarroll, C. N. Pato, M. T. Pato, N. Craddock, M. Landen, J. W. Smoller, and P. Sklar. Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder. *Transl Psychiatry*, 7(1):e993, 2017.

[46] J. Chen, R. Swofford, J. Johnson, B. B. Cummings, N. Rogel, K. Lindblad-Toh, W. Haerty, F. D. Palma, and A. Regev. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res*, 29(1):53–63, 2019.

[47] Jenny Chen, Ross Swofford, Jeremy Johnson, Beryl B. Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica di Palma, and Aviv Regev. A quantitative model for characterizing the evolutionary history of mammalian gene expression. *bioRxiv*, 2017.

[48] 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.

[49] B. P. Chumpitazi, J. L. Cope, E. B. Hollister, C. M. Tsai, A. R. McMeans, R. A. Luna, J. Versalovic, and R. J. Shulman. Randomised clinical trial: gut microbiome biomarkers are associated with clinical response to a low fodmap diet in children with the irritable bowel syndrome. *Aliment Pharmacol Ther*, 42(4):418–27, 2015.

[50] 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.

[51] R. C. Colligan, R. J. Ferdinande, A. R. Lucas, and J. W. Duncan. A one-year followup study of adolescent patients hospitalized with anorexia nervosa. *J Dev Behav Pediatr*, 4(4):278–9, 1983.

[52] M. A. Crisafulli, A. Von Holle, and C. M. Bulik. Attitudes towards anorexia nervosa: the impact of framing on blame and stigma. *Int J Eat Disord*, 41(4):333–9, 2008.

[53] A. H. Crisp, K. R. Norton, S. Jurczak, C. Bowyer, and S. Duncan. A treatment approach to anorexia nervosa–25 years on. *J Psychiatr Res*, 19(2-3):393–404, 1985.

[54] R. Dalle Grave, M. Sartirana, M. El Ghoch, and S. Calugi. Dsm-5 severity specifiers for anorexia nervosa and treatment outcomes in adult females. *Eat Behav*, 31:18–23, 2018.

[55] 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.

[56] E. R. Davenport, J. G. Sanders, S. J. Song, K. R. Amato, A. G. Clark, and R. Knight. The human microbiome in evolution. *BMC Biol*, 15(1):127, 2017.

[57] K. Day, L. L. Waite, A. Thalacker-Mercer, A. West, M. M. Bamman, J. D. Brooks, R. M. Myers, and D. Absher. Differential dna methylation with age displays both common and dynamic features across human tissues that are influenced by cpg landscape. *Genome Biol*, 14(9):R102, 2013.

[58] A. Deep-Soboslay, F. M. Benes, V. Haroutunian, J. K. Ellis, J. E. Kleinman, and T. M. Hyde. Psychiatric brain banking: three perspectives on current trends and future directions. *Biol Psychiatry*, 69(2):104–12, 2011.

[59] Z. Dezso, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. J. Brennan, A. Guryanov, K. Li, J. Blake, R. R. Samaha, and T. Nikolskaya. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol*, 6:49, 2008.

[60] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[61] A. Dobin and T. R. Gingeras. Mapping rna-seq reads with star. *Curr Protoc Bioinformatics*, 51:11 14 1–19, 2015.

[62] L. Duncan, Z. Yilmaz, H. Gaspar, R. Walters, J. Goldstein, V. Anttila, B. Bulik-Sullivan, S. Ripke, Consortium Eating Disorders Working Group of the Psychiatric Genomics, L. Thornton, A. Hinney, M. Daly, P. F. Sullivan, E. Zeggini, G. Breen, and C. M. Bulik. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *Am J Psychiatry*, 174(9):850–858, 2017.

[63] L. E. Duncan, A. Ratanatharathorn, A. E. Aiello, L. M. Almli, A. B. Amstadter, A. E. Ashley-Koch, D. G. Baker, J. C. Beckham, L. J. Bierut, J. Bisson, B. Bradley, C. Y. Chen, S. Dalvie, L. A. Farrer, S. Galea, M. E. Garrett, J. E. Gelernter, G. Guffanti, M. A. Hauser, E. O. Johnson, R. C. Kessler, N. A. Kimbrel, A. King, N. Koen, H. R. Kranzler, M. W. Logue, A. X. Maihofer, A. R. Martin, M. W. Miller, R. A. Morey, N. R. Nugent, J. P. Rice, S. Ripke, A. L. Roberts, N. L. Saccone, J. W. Smoller, D. J. Stein, M. B. Stein, J. A. Sumner, M. Uddin, R. J. Ursano, D. E. Wildman, R. Yehuda, H. Zhao, M. J. Daly, I. Liberzon, K. J. Ressler, C. M. Nievergelt, and K. C. Koenen. Largest gwas of ptsd (n=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. *Mol Psychiatry*, 23(3):666–673, 2018.

[64] K. Bore E, C. Apostel, S. Halicki, Y. Kuzyakov, and M. A. Dippold. Soil microorganisms can overcome respiration inhibition by coupling intra- and extracellular metabolism: (13)c metabolic tracing reveals the mechanisms. *ISME J*, 11(6):1423–1433, 2017.

[65] M. M. Easter. "not all my fault": genetics, stigma, and personal responsibility for women with eating disorders. *Soc Sci Med*, 75(8):1408–16, 2012.

[66] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck. Dna methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38(12):1378–85, 2006.

[67] K. T. Eddy, N. Tabri, J. J. Thomas, H. B. Murray, A. Keshaviah, E. Hastings, K. Edkins, M. Krishna, D. B. Herzog, P. K. Keel, and D. L. Franko. Recovery from anorexia nervosa and bulimia nervosa at 22-year follow-up. *J Clin Psychiatry*, 78(2):184–189, 2017.

[68] R. Edgar, M Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. 30:207–210, 2002.

[69] J. T. Eppig, C. L. Smith, J. A. Blake, M. Ringwald, J. A. Kadin, J. E. Richardson, and C. J. Bult. Mouse genome informatics (mgi): Resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. *Methods Mol Biol*, 1488:47–73, 2017.

[70] Jason Ernst and Ziv Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics*, 7(1):191, 2006.

[71] 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.

[72] A. K. Fairweather-Schmidt and T. D. Wade. Dsm-5 eating disorders and other specified eating and feeding disorders: is there a meaningful differentiation? *Int J Eat Disord*, 47(5):524–33, 2014.

[73] M. M. Fichter, N. Quadflieg, R. D. Crosby, and S. Koch. Long-term outcome of anorexia nervosa: Results from a large clinical longitudinal study. *Int J Eat Disord*, 50(9):1018–1030, 2017.

[74] D. L. Finfgeld. Anorexia nervosa: Analysis of long-term outcomes and clinical implications. *Archives of Psychiatric Nursing*, 16(4):176–186, 2002.

[75] S. F. Forman, N. McKenzie, R. Hehn, M. C. Monge, C. J. Kapphahn, K. A. Mammel, S. T. Callahan, E. J. Sigel, T. Bravender, M. Romano, E. S. Rome, K. A. Robinson, M. Fisher, J. B. Malizio, D. S. Rosen, A. C. Hergenroeder, S. M. Buckelew, M. S. Jay, J. Lindenbaum, V. I. Rickert, A. Garber, N. H. Golden, and E. R. Woods. Predictors of outcome at 1 year in adolescents with dsm-5 restrictive eating disorders: report of the national eating disorders quality improvement collaborative. *J Adolesc Health*, 55(6):750–6, 2014.

[76] A. Fortna, Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, A. Karimpour-Fard, D. Glueck, L. McGavran, R. Berry, J. Pollack, and J. M. Sikela. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol*, 2(7):E207, 2004.

[77] L. C. Fosburgh. Mack: an autobiography. *AANA J*, 64(6):583–6, 1996.

[78] D. L. Franko, A. Keshaviah, K. T. Eddy, M. Krishna, M. C. Davis, P. K. Keel, and D. B. Herzog. A longitudinal investigation of mortality in anorexia nervosa and bulimia nervosa. *Am J Psychiatry*, 170(8):917–25, 2013.

[79] D. L. Franko, N. Tabri, A. Keshaviah, H. B. Murray, D. B. Herzog, J. J. Thomas, K. Coniglio, P. K. Keel, and K. T. Eddy. Predictors of long-term recovery in anorexia nervosa and bulimia nervosa: Data from a 22-year longitudinal study. *J Psychiatr Res*, 96:183–188, 2018.

[80] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.

[81] 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. doi: 10.7554/eLife.07103, 2015.

[82] 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.

[83] P. R. Gibson, E. P. Halmos, and J. G. Muir. Letter: bias in clinical trials of the symptomatic effects of the low fodmap diet for irritable bowel syndrome-getting the facts right. *Aliment Pharmacol Ther*, 46(3):385–386, 2017.

[84] 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.

[85] M. Gjoertz, J. Wang, S. Chatelet, C. Monney Chaubert, F. Lier, and A. E. Ambresin. Nutrition approach for inpatients with anorexia nervosa: Impact of a clinical refeeding guideline. *JPEN J Parenter Enteral Nutr*, 2019.

[86] Jr. Goff, D. C., D. M. Lloyd-Jones, G. Bennett, S. Coady, Sr. D'Agostino, R. B., R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson,

J. S. Schwartz, S. T. Shero, Jr. Smith, S. C., P. Sorlie, N. J. Stone, and P. W. F. Wilson. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *J Am Coll Cardiol*, 63(25 Pt B):2935–2959, 2014.

[87] Data Analysis GTEx Consortium, Group Coordinating Center Analysis Working, Group Statistical Methods groups Analysis Working, GTEx groups Enhancing, N. I. H. Common Fund, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, Ndri Biospecimen Collection Source Site, Rpci Biospecimen Collection Source Site, Vari Biospecimen Core Resource, Bank Brain Bank Repository-University of Miami Brain Endowment, Management Leidos Biomedical-Project, Elsi Study, Integration Genome Browser Data, E. B. I. Visualization, Integration Genome Browser Data, University of California Santa Cruz Visualization-Ucsc Genomics Institute, analysts Lead, Data Analysis Laboratory, Center Coordinating, N. I. H. program management, collection Biospecimen, Pathology, Q. T. L. manuscript working group e, A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[88] K. D. Hansen, B. Langmead, and R. A. Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10):R83, 2012.

[89] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22(9):1760–74, 2012.

[90] A. H. Hart, L. Hartley, K. Sourris, E. S. Stadler, R. Li, E. G. Stanley, P. P. Tam, A. G. Elefanty, and L. Robb. Mixl1 is required for axial mesendoderm morphogenesis and patterning in the murine embryo. *Development*, 129(15):3597–608, 2002.

[91] I. Hernando-Herraez, R. Garcia-Perez, A. J. Sharp, and T. Marques-Bonet. Dna methylation: Insights into human evolution. *PLoS Genet*, 11(12):e1005661, 2015.

[92] I. Hernando-Herraez, H. Heyn, M. Fernandez-Callejo, E. Vidal, H. Fernandez-Bellon, J. Prado-Martinez, A. J. Sharp, M. Esteller, and T. Marques-Bonet. The interplay between dna methylation and sequence divergence in recent human evolution. *Nucleic Acids Res*, 43(17):8204–14, 2015.

[93] 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.

[94] D. B. Herzog, N. R. Sacks, M. B. Keller, P. W. Lavori, K. B. von Ranson, and H. M. Gray. Patterns and predictors of recovery in anorexia nervosa and bulimia nervosa. *J Am Acad Child Adolesc Psychiatry*, 32(4):835–42, 1993.

[95] I. Hetman, A. Brunstein Klomek, G. Goldzweig, A. Hadas, M. Horwitz, and S. Fennig. Percentage from target weight (pftw) predicts re-hospitalization in adolescent anorexia nervosa. *Isr J Psychiatry Relat Sci*, 54(3):28–34, 2017.

[96] C. Hippman, Z. Lohn, A. Ringrose, A. Inglis, J. Cheek, and J. C. Austin. "nothing is absolute in life": understanding uncertainty in the context of psychiatric genetic counseling from the perspective of those with serious mental illness. *J Genet Couns*, 22(5):625–32, 2013.

[97] H. W. Hoek. Classification, epidemiology and treatment of dsm-5 feeding and eating disorders. *Curr Opin Psychiatry*, 26(6):529–31, 2013.

[98] K. Holtkamp, J. Hebebrand, C. Mika, M. Heer, N. Heussen, and B. Herpertz-Dahlmann. High serum leptin levels subsequent to weight gain predict renewed weight loss in patients with anorexia nervosa. *Psychoneuroendocrinology*, 29(6):791–7, 2004.

[99] H. Hori, D. Sasayama, T. Teraishi, N. Yamamoto, S. Nakamura, M. Ota, K. Hattori, Y. Kim, T. Higuchi, and H. Kunugi. Blood-based gene expression signatures of medication-free outpatients with major depressive disorder: integrative genome-wide and candidate gene analyses. *Sci Rep*, 6:18776, 2016.

[100] S. Horvath. Dna methylation age of human tissues and cell types. *Genome Biol*, 14(10):R115, 2013.

[101] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics Proteomics*, 15(1):41–51, 2018.

[102] C. Hubel, H. A. Gaspar, J. R. I. Coleman, H. Finucane, K. L. Purves, K. B. Hanscombe, I. Prokopenko, Magic investigators, M. Graff, J. S. Ngwa, T. Workalemahu, Consortium Eating Disorders Working Group of the Psychiatric Genomics, Consortium Major Depressive Disorder Working Group of the Psychiatric Genomics, Consortium Schizophrenia Working Group of the Psychiatric Genomics, Consortium Tourette Syndrome/Obsessive-Compulsive Disorder Working Group of the Psychiatric Genomics, P. F. O'Reilly, C. M. Bulik, and G. Breen. Genomics of body fat percentage may contribute to sex bias in anorexia nervosa. *Am J Med Genet B Neuropsychiatr Genet*, 180(6):428–438, 2019.

[103] C. Hubel, S. J. Marzi, G. Breen, and C. M. Bulik. Epigenetics in eating disorders: a systematic review. *Mol Psychiatry*, 24(6):901–915, 2019.

[104] G. Jagielska and I. Kacperska. Outcome, comorbidity and prognosis in anorexia nervosa. *Psychiatr Pol*, 51(2):205–218, 2017.

[105] G. Jagielska, J. Przedlacki, Z. Bartoszewicz, A. Kondracka, A. Butwicka, E. Racicka, O. Rowinski, R. Brzewski, T. Wolanczyk, C. Tomaszewicz-Libudzic, and U. Szymanska. Bone mineralization and densitometric evaluation of vertebral fractures in women 6-21 years after the onset of anorexia nervosa symptoms. *Psychiatr Pol*, 51(2):231–246, 2017.

[106] C. Ju, L. M. Fiori, R. Belzeaux, J. F. Theroux, G. G. Chen, Z. Aouabed, P. Blier, F. Farzan, B. N. Frey, P. Giacobbe, R. W. Lam, F. Leri, G. M. MacQueen, R. Milev, D. J. Muller, S. V. Parikh, S. Rotzinger, C. N. Soares, R. Uher, Q. Li, J. A. Foster, S. H. Kennedy, and G. Turecki. Integrated genome-wide methylation and expression analyses reveal functional predictors of response to antidepressants. *Transl Psychiatry*, 9(1):254, 2019.

[107] A. S. Kaplan, B. T. Walsh, M. Olmsted, E. Attia, J. C. Carter, M. J. Devlin, K. M. Pike, B. Woodside, W. Rockert, C. A. Roberto, and M. Parides. The slippery slope: prediction of successful weight maintenance in anorexia nervosa. *Psychol Med*, 39(6):1037–45, 2009.

[108] M. W. Karaman, M. L. Houck, L. G. Chemnick, S. Nagpal, D. Chawannakul, D. Sudano, B. L. Pike, V. V. Ho, O. A. Ryder, and J. G. Hacia. Comparative analysis of gene-expression patterns in human and african great ape cultured fibroblasts. *Genome Res*, 13(7):1619–30, 2003.

[109] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Res*, 42(Database issue):D764–70, 2014.

[110] SW Kembel, PD Cowan, MR Helmus, WK Cornwell, H Morlon, DD Ackerly, SP Blomberg, and CO Webb. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26(11):1463–1464, 2010.

[111] P. Khaitovich, I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, 309(5742):1850–4, 2005.

[112] P. Khaitovich, B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H. H. Do, G. Weiss, W. Enard, F. Heissig, T. Arendt, K. Nieselt-Struwe, E. E. Eichler, and S. Paabo. Regional patterns of gene expression in human and chimpanzee brains. *Genome Res*, 14(8):1462–73, 2004.

[113] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. A neutral model of transcriptome evolution. *PLoS Biol*, 2(5):E132, 2004.

[114] S. S. Khalsa, L. C. Portnoff, D. McCurdy-McKinnon, and J. D. Feusner. What happens after treatment? a systematic review of relapse, remission, and recovery in anorexia nervosa. *J Eat Disord*, 5:20, 2017.

[115] Z. Khan, M. J. Ford, D. A. Cusanovich, A. Mitrano, J. K. Pritchard, and Y. Gilad. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, 342(6162):1100–4, 2013.

[116] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4), 2013.

[117] Y. Kim, S. E. Trace, J. J. Crowley, K. A. Brownley, R. M. Hamer, D. S. Pisetsky, P. F. Sullivan, and C. M. Bulik. Assessment of gene expression in peripheral blood using rnaseq before and after weight restoration in anorexia nervosa. *Psychiatry Res*, 210(1):287–93, 2013.

[118] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–16, 1975.

[119] G. S. Kuipers, S. D. Hollander, L. A. van der Ark, and M. H. J. Bekker. Recovery from eating disorder 1 year after start of treatment is related to better mentalization and strong reduction of sensitivity to others. *Eat Weight Disord*, 22(3):535–547, 2017.

[120] P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[121] 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.

[122] B. Lemos, C. D. Meiklejohn, M. Caceres, and D. L. Hartl. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution*, 59(1):126–37, 2005.

[123] 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.

[124] 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.

[125] C. Lindskog. The human protein atlas - an important resource for basic and clinical research. *Expert Rev Proteomics*, 13(7):627–9, 2016.

[126] J. Lock, W. S. Agras, D. Le Grange, J. Couturier, D. Safer, and S. W. Bryson. Do end of treatment assessments predict outcome at follow-up in eating disorders? *International Journal of Eating Disorders*, 46(8):771–778, 2013.

[127] M. Logotheti, O. Papadodima, N. Venizelos, A. Chatziioannou, and F. Kolisis. A comparative genomic study in schizophrenic and in bipolar disorder patients, based on microarray expression profiling meta-analysis. *ScientificWorldJournal*, 2013:685917, 2013.

[128] 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.

[129] P. P. Machado, C. M. Grilo, and R. D. Crosby. Evaluation of the dsm-5 severity indicator for anorexia nervosa. *Eur Eat Disord Rev*, 25(3):221–223, 2017.

[130] S. Magnani, S. Roberto, G. Sainas, R. Milia, G. Palazzolo, L. Cugusi, V. Pinna, A. Doneddu, S. A. H. Kakhak, F. Tocco, G. Mercuro, and A. Crisafulli. Metaboreflex-mediated hemodynamic abnormalities in individuals with coronary artery disease without overt signs or symptoms of heart failure. *Am J Physiol Heart Circ Physiol*, 314(3):H452–H463, 2018.

[131] Gwas Consortium Major Depressive Disorder Working Group of the Psychiatric, S. Ripke, N. R. Wray, C. M. Lewis, S. P. Hamilton, M. M. Weissman, G. Breen, E. M. Byrne, D. H. Blackwood, D. I. Boomsma, S. Cichon, A. C. Heath, F. Holsboer, S. Lucae, P. A. Madden, N. G. Martin, P. McGuffin, P. Muglia, M. M. Noethen, B. P. Penninx, M. L. Pergadia, J. B. Potash, M. Rietschel, D. Lin, B. Muller-Myhsok, J. Shi, S. Steinberg, H. J. Grabe, P. Lichtenstein, P. Magnusson, R. H. Perlis, M. Preisig, J. W. Smoller, K. Stefansson, R. Uher, Z. Kutalik, K. E. Tansey, A. Teumer, A. Viktorin, M. R. Barnes, T. Bettecken, E. B. Binder, R. Breuer, V. M. Castro, S. E. Churchill, W. H. Coryell, N. Craddock, I. W. Craig, D. Czamara, E. J. De Geus, F. Degenhardt, A. E. Farmer, M. Fava, J. Frank, V. S. Gainer, P. J. Gallagher, S. D. Gordon, S. Goryachev, M. Gross, M. Guipponi, A. K. Henders, S. Herms, I. B. Hickie, S. Hoefels, W. Hoogendijk, J. J. Hottenga, D. V. Iosifescu, M. Ising, I. Jones, L. Jones, T. Jung-Ying, J. A. Knowles, I. S. Kohane, M. A. Kohli, A. Korszun, M. Landen, W. B. Lawson, G. Lewis, D. Macintyre, W. Maier, M. Mattheisen, P. J. McGrath, A. McIntosh, A. McLean, C. M. Middeldorp, L. Middleton, G. M. Montgomery, S. N. Murphy, M. Nauck, W. A. Nolen, D. R. Nyholt, M. O'Donovan, H. Oskarsson, N. Pedersen, W. A. Scheftner, A. Schulz, T. G. Schulze, S. I. Shyn, E. Sigurdsson, S. L. Slager, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry*, 18(4):497–511, 2013.

[132] S. H. Makhzoumi, J. W. Coughlin, C. C. Schreyer, G. W. Redgrave, S. C. Pitts, and A. S. Guarda. Weight gain trajectories in hospital-based treatment of anorexia nervosa. *Int J Eat Disord*, 50(3):266–274, 2017.

[133] 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.

[134] D. I. Martin, M. Singer, J. Dhahbi, G. Mao, L. Zhang, G. P. Schroth, L. Pachter, and D. Boffelli. Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res*, 21(12):2049–57, 2011.

[135] 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.

[136] L. E. Mayer, C. A. Roberto, D. R. Glasofer, S. F. Etu, D. Gallagher, J. Wang, S. B. Heymsfield, Jr. Pierson, R. N., E. Attia, M. J. Devlin, and B. T. Walsh. Does percent body fat predict outcome in anorexia nervosa? *Am J Psychiatry*, 164(6):970–2, 2007.

[137] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res*, 20(9):1297–303, 2010.

[138] E. Mekori, L. Halevy, S. I. Ziv, A. Moreno, A. Enoch-Levy, A. Weizman, and D. Stein. Predictors of short-term outcome variables in hospitalised female adolescents with eating disorders. *International Journal of Psychiatry in Clinical Practice*, 21(1):41–49, 2017.

[139] A. Menke. Gene expression: biomarker of antidepressant therapy? *Int Rev Psychiatry*, 25(5):579–91, 2013.

[140] J. Merkin, C. Russell, P. Chen, and C. B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–9, 2012.

[141] M. Misra, R. Prabhakaran, K. K. Miller, M. A. Goldstein, D. Mickley, L. Clauss, P. Lockhart, J. Cord, D. B. Herzog, D. K. Katzman, and A. Klibanski. Weight gain and restoration of menses as predictors of bone mineral density change in adolescent girls with anorexia nervosa-1. *Journal of Clinical Endocrinology and Metabolism*, 93(4):1231–1237, 2008.

[142] A. Molaro, E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon, and A. D. Smith. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, 146(6):1029–41, 2011.

[143] P. Monteleone, M. Di Genio, A. M. Monteleone, C. Di Filippo, and M. Maj. Investigation of factors associated to crossover from anorexia nervosa restricting type (anr) and anorexia nervosa binge-purging type (anbp) to bulimia nervosa and comparison of bulimia nervosa patients with or without previous anr or anbp. *Compr Psychiatry*, 52(1):56–62, 2011.

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

[145] M. A. Munn-Chernoff and J. H. Baker. A primer on the genetics of comorbid eating disorders and substance use disorders. *Eur Eat Disord Rev*, 24(2):91–100, 2016.

[146] Venables W. N. and Ripley B. D. *Modern Applied Statistics with S*. 2002.

[147] Y. Nakai, K. Nin, S. Noma, S. Teramukai, K. Fujikawa, and S. A. Wonderlich. The impact of dsm-5 on the diagnosis and severity indicator of eating disorders in a treatment-seeking sample. *Int J Eat Disord*, 50(11):1247–1254, 2017.

[148] B. M. Neale, S. E. Medland, S. Ripke, P. Asherson, B. Franke, K. P. Lesch, S. V. Faraone, T. T. Nguyen, H. Schafer, P. Holmans, M. Daly, H. C. Steinhausen, C. Freitag, A. Reif, T. J. Renner, M. Romanos, J. Romanos, S. Walitza, A. Warnke, J. Meyer, H. Palmason, J. Buitelaar, A. A. Vasquez, N. Lambregts-Rommelse, M. Gill, R. J. Anney, K. Langely, M. O'Donovan, N. Williams, M. Owen, A. Thapar, L. Kent, J. Sergeant, H. Roeyers, E. Mick, J. Biederman, A. Doyle, S. Smalley, S. Loo, H. Hakonarson, J. Elia, A. Todorov, A. Miranda, F. Mulas, R. P. Ebstein, A. Rothenberger, T. Banaschewski, R. D. Oades, E. Sonuga-Barke, J. McGough, L. Nisenbaum, F. Middleton, X. Hu, S. Nelson, and Gwas Consortium Adhd Subgroup Psychiatric. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*, 49(9):884–97, 2010.

[149] P. D. Negraes, F. R. Cugola, R. H. Herai, C. A. Trujillo, A. S. Cristino, T. Chailangkarn, A. R. Muotri, and V. Duvvuri. Modeling anorexia nervosa: transcriptional insights from human ipsc-derived neurons. *Transl Psychiatry*, 7(3):e1060, 2017.

[150] J. P. Nogueira, R. Valero, M. Maraninchi, A. M. Lorec, C. Samuelian-Massat, A. Begu-Le Corroller, A. Nicolay, J. Gaudart, H. Portugal, and B. Vialettes. Growth hormone level at admission and its evolution during refeeding are predictive of short-term outcome in restrictive anorexia nervosa. *British Journal of Nutrition*, 109(12):2175–2181, 2013.

[151] 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.

[152] X Nuttle, G Giannuzzi, MH Duyzend, JG Schraiber, I Narvaiza, PH Sudmant, O Penn, G Chiatante, M Malig, J Huddleston, C Benner, F Camponeschi, S Ciofi-Baffoni, HAF Stessman, MCN Marchetto, L Denman, L Harshman, C Baker, A Raja, K Penewit, N Janke, WJ Tang, M Ventura, L Banci, F Antonacci, JM Akey, CT Amemiya, FH Gage, A Reymond, and EE Eichler. Emergence of a homo sapiens-specific gene family and chromosome 16p11.2 cnv susceptibility. *Nature*, 536(7615):205–209, 2016.

[153] A. A. Pai, J. T. Bell, J. C. Marioni, J. K. Pritchard, and Y. Gilad. A genome-wide study of dna methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet*, 7(2):e1001316, 2011.

[154] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*, 14(4):417–419, 2017.

[155] G. H. Perry, P. Melsted, J. C. Marioni, Y. Wang, R. Bainer, J. K. Pickrell, K. Michelini, S. Zehr, A. D. Yoder, M. Stephens, J. K. Pritchard, and Y. Gilad. Comparative rna sequencing reveals substantial genetic variation in endangered primates. *Genome Res*, 22(4):602–10, 2012.

[156] E. Pierson, G. TEx Consortium, D. Koller, A. Battle, S. Mostafavi, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput Biol*, 11(5):e1004220, 2015.

[157] J. Prado-Martinez, P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, A. Cagan, C. Theunert, F. Casals, H. Laayouni, K. Munch, A. Hobolth, A. E. Halager, M. Malig, J. Hernandez-Rodriguez, I. Hernando-Herraez, K. Prufer, M. Pybus, L. Johnstone, M. Lachmann, C. Alkan, D. Twigg, N. Petit, C. Baker, F. Hormozdiari, M. Fernandez-Callejo, M. Dabad, M. L. Wilson, L. Stevison, C. Camprubi, T. Carvalho, A. Ruiz-Herrera, L. Vives, M. Mele, T. Abello, I. Kondova, R. E. Bontrop, A. Pusey, F. Lankester, J. A. Kiyang, R. A. Bergl, E. Lonsdorf, S. Myers, M. Ventura, P. Gagneux, D. Comas, H. Siegismund, J. Blanc, L. Agueda-Calpena, M. Gut, L. Fulton, S. A. Tishkoff, J. C. Mullikin, R. K. Wilson, I. G. Gut, M. K. Gonder, O. A. Ryder, B. H. Hahn, A. Navarro, J. M. Akey, J. Bertranpetit, D. Reich, T. Mailund, M. H. Schierup, C. Hvilsom, A. M. Andres, J. D. Wall, C. D. Bustamante, M. F. Hammer, E. E. Eichler, and T. Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–5, 2013.

[158] SL Prescott, R Srinivasan, MC Marchetto, I Grishina, I Narvaiza, L Selleri, FH 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.

[159] Gwas Consortium Bipolar Disorder Working Group Psychiatric. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near odz4. *Nat Genet*, 43(10):977–83, 2011.

[160] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[161] J. D. Rainer. Genetic knowledge and heredity counseling: new responsibilities for psychiatry. *Proc Annu Meet Am Psychopathol Assoc*, (63):289–95, 1975.

[162] 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.

[163] D. Regier. Interview with darrel a. regier. the developmental process for the diagnostic and statistical manual of mental disorders, fifth edition. interview by norman sussman. *CNS Spectr*, 13(2):120–4, 2008.

[164] ME Ritchie, B Phipson, D Wu, YF Hu, CW Law, W Shi, and GK Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47. doi:10.1093/nar/gkv007, 2015.

[165] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

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

[167] Irene Gallego Romero, Ilya Ruvinsky, and Yoav Gilad. Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516, 2012.

[168] F. M. Rosa. Mix.1, a homeobox mrna inducible by mesoderm inducers, is expressed mostly in the presumptive endodermal cells of xenopus embryos. *Cell*, 57(6):965–74, 1989.

[169] A. P. Russ, S. Wattler, W. H. Colledge, S. A. Aparicio, M. B. Carlton, J. J. Pearce, S. C. Barton, M. A. Surani, K. Ryan, M. C. Nehls, V. Wilson, and M. J. Evans. Eomesodermin is required for mouse trophoblast development and mesoderm formation. *Nature*, 404(6773):95–9, 2000.

[170] J. Ryan, A. Virani, and J. C. Austin. Ethical issues associated with genetic counseling in the context of adolescent psychiatry. *Appl Transl Genom*, 5:23–9, 2015.

[171] K. G. Sagaser, S. Shahrukh Hashmi, R. D. Carter, J. Lemons, H. Mendez-Figueroa, S. Nassef, B. Peery, and C. N. Singletary. Spiritual exploration in the prenatal genetic counseling session. *J Genet Couns*, 25(5):923–35, 2016.

[172] A. Saha, Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, G. TEx Consortium, B. E. Engelhardt, and A. Battle. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res*, 27(11):1843–1858, 2017.

[173] D. L. Santos Ferreira, C. Hubel, M. Herle, M. Abdulkadir, R. J. F. Loos, R. Bryant-Waugh, C. M. Bulik, B. L. De Stavola, D. A. Lawlor, and N. Micali. Associations between blood metabolic profile at 7 years old and eating disorders in adolescence: Findings from the avon longitudinal study of parents and children. *Metabolites*, 9(9), 2019.

[174] K. Schaumberg, E. Welch, L. Breithaupt, C. Hubel, J. H. Baker, M. A. Munn-Chernoff, Z. Yilmaz, S. Ehrlich, L. Mustelin, A. Ghaderi, A. J. Hardaway, E. C. Bulik-Sullivan, A. M. Hedman, A. Jangmo, I. A. K. Nilsson, C. Wiklund, S. Yao, M. Seidel, and C. M. Bulik. The science behind the academy for eating disorders' nine truths about eating disorders. *Eur Eat Disord Rev*, 25(6):432–450, 2017.

[175] U. M. Schulze, S. Schuler, D. Schlamp, P. Schneider, and C. Mehler-Wex. Bone mineral density in partially recovered early onset anorexic patients - a follow-up investigation. *Child Adolesc Psychiatry Ment Health*, 4:20, 2010.

[176] A. J. Sharp, E. Stathaki, E. Migliavacca, M. Brahmachary, S. B. Montgomery, Y. Dupre, and S. E. Antonarakis. Dna methylation profiles of human active and inactive x chromosomes. *Genome Res*, 21(10):1592–600, 2011.

[177] Y. Shibata, N. C. Sheffield, O. Fedrigo, C. C. Babbitt, M. Wortham, A. K. Tewari, D. London, L. Song, B. K. Lee, V. R. Iyer, S. C. Parker, E. H. Margulies, G. A. Wray, T. S. Furey, and G. E. Crawford. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet*, 8(6):e1002789, 2012.

[178] ML Siegal and A Bergman. Waddington's canalization revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences*, 99(16):10528–10532, 2002.

[179] C. L. Smith, Blake J.A., J. A. Kadin, Richardson J.E., C. J. Bult, and the Mouse Genome Database Group. Mouse genome database (mgd)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res*, 46:D836–D842, 2018.

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

[181] Gordon K. Smyth, Walter, Australia Eliza Hall Institute of Medical Research Melbourne, Vic 3050, Jolle Michaud, Walter, Australia Eliza Hall Institute of Medical Research Melbourne, Vic 3050, Hamish S. Scott, Walter, and Australia Eliza Hall Institute of Medical Research Melbourne, Vic 3050. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

[182] K. Soddy. Maternal care and mental health. *Ment Health (Lond)*, 11(2):70–75, 1952.

[183] 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.

[184] C. Soneson, M. I. Love, and M. D. Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Res*, 4:1521, 2015.

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

[186] H. C. Steinhausen, M. Grigoroiu-Serbanescu, S. Boyadjieva, K. J. Neumarker, and C. Winkler Metzke. Course and predictors of rehospitalization in adolescent anorexia nervosa in a multisite study. *Int J Eat Disord*, 41(1):29–36, 2008.

[187] Isaac Stephens. *The gentlewoman's remembrance : patriarchy, piety, and singlehood in early Stuart England*. 2016.

[188] Jr. Stephens, Jonathan. *Prostitutes in the pews*. 2016.

[189] A. B. Stergachis, S. Neph, R. Sandstrom, E. Haugen, A. P. Reynolds, M. Zhang, R. Byron, T. Canfield, S. Stelhing-Sun, K. Lee, R. E. Thurman, S. Vong, D. Bates, F. Neri, M. Diegel, E. Giste, D. Dunn, J. Vierstra, R. S. Hansen, A. K. Johnson, P. J. Sabo, M. S. Wilken, T. A. Reh, P. M. Treuting, R. Kaul, M. Groudine, M. A. Bender, E. Borenstein, and J. A. Stamatoyannopoulos. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, 515(7527):365–70, 2014.

[190] S. Stojiljkovic-Drobnjak, S. Fischer, M. Arnold, W. Langhans, and U. Ehlert. Menopause is associated with decreased postprandial ghrelin, whereas a history of anorexia nervosa is associated with increased total ghrelin. *J Neuroendocrinol*, 31(7):e12661, 2019.

[191] 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.

[192] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[193] E. Stroe-Kunold, M. Buckert, H. C. Friederich, D. Wesche, S. Kopf, W. Herzog, and B. Wild. Time course of leptin in patients with anorexia nervosa during inpatient treatment: Longitudinal relationships to bmi and psychological factors. *Plos One*, 11(12), 2016.

[194] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, 2003.

[195] P. F. Sullivan, A. Agrawal, C. M. Bulik, O. A. Andreassen, A. D. Borglum, G. Breen, S. Cichon, H. J. Edenberg, S. V. Faraone, J. Gelernter, C. A. Mathews, C. M. Nievergelt, J. W. Smoller, M. C. O'Donovan, and Consortium Psychiatric Genomics. Psychiatric genomics: An update and an agenda. *Am J Psychiatry*, 175(1):15–27, 2018.

[196] F. Supek, M. Bosnjak, N. Skunca, and T. Smuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, 2011.

[197] 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.

[198] The Encode Project. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[199] D. M. Thomas, J. E. Navarro-Barrientos, D. E. Rivera, S. B. Heymsfield, C. Bredlau, L. M. Redman, C. K. Martin, S. A. Lederman, M. Collins L, and N. F. Butte. Dynamic energy-balance model predicting gestational weight gain. *Am J Clin Nutr*, 95(1):115–22, 2012.

[200] PD Thomas, A Kejariwal, N Guo, HY Mi, MJ Campbell, A Muruganujan, and B Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mrna/protein expression analysis and coding snp scoring tools. *Nucleic Acids Research*, 34:W645–W650, 2006.

[201] P. J. Thul and C. Lindskog. The human protein atlas: A spatial map of the human proteome. *Protein Sci*, 27(1):233–244, 2018.

[202] A. M. Tsankov, H. Gu, V. Akopian, M. J. Ziller, J. Donaghey, I. Amit, A. Gnirke, and A. Meissner. Transcription factor binding dynamics during human es cell differentiation. *Nature*, 518(7539):344–9, 2015.

[203] J. Tsuji and Z. Weng. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. *Brief Bioinform*, 17(6):938–952, 2016.

[204] 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.

[205] B van de Geijn, G McVicker, Y Gila, and JK Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11):1061–1063, 2015.

[206] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*, 43:11 10 1–33, 2013.

[207] A. A. van Elburg, M. J. Kas, J. J. Hillebrand, R. J. Eijkemans, and H. van Engeland. The impact of hyperactivity and leptin on recovery from anorexia nervosa. *J Neural Transm (Vienna)*, 114(9):1233–7, 2007.

[208] D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, and D. T. Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–66, 2015.

[209] 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.

[210] I. Virgolini. Mack forster award lecture. receptor nuclear medicine: vasointestinal peptide and somatostatin receptor scintigraphy for diagnosis and treatment of tumour patients. *Eur J Clin Invest*, 27(10):793–800, 1997.

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

[212] CH Waddington. Canalization of development and genetic assimilation of acquired characters. *Nature*, 183:1654–1655, 1959.

[213] M. C. Ward, M. D. Wilson, N. L. Barbosa-Morais, D. Schmidt, R. Stark, Q. Pan, P. C. Schwalie, S. Menon, M. Lukk, S. Watt, D. Thybert, C. Kutter, K. Kirschner, P. Flicek, B. J. Blencowe, and D. T. Odom. Latent regulatory potential of human-specific repetitive elements. *Mol Cell*, 49(2):262–72, 2013.

[214] M. Warnefors and A. Eyre-Walker. A selection index for gene expression evolution and its application to the divergence between humans and chimpanzees. *PLoS One*, 7(4):e34935, 2012.

[215] H. J. Watson, Z. Yilmaz, L. M. Thornton, C. Hubel, J. R. I. Coleman, H. A. Gaspar, J. Bryois, A. Hinney, V. M. Leppa, M. Mattheisen, S. E. Medland, S. Ripke, S. Yao, P. Giusti-Rodriguez, Initiative Anorexia Nervosa Genetics, K. B. Hanscombe, K. L. Purves, Consortium Eating Disorders Working Group of the Psychiatric Genomics, R. A. H. Adan, L. Alfredsson, T. Ando, O. A. Andreassen, J. H. Baker, W. H. Berrettini, I. Boehm, C. Boni, V. B. Perica, K. Buehren, R. Burghardt, M. Cassina, S. Cichon, M. Clementi, R. D. Cone, P. Courtet, S. Crow, J. J. Crowley, U. N. Danner, O. S. P. Davis, M. de Zwaan, G. Dedoussis, D. Degortes, J. E. DeSocio, D. M. Dick, D. Dikeos, C. Dina, M. Dmitrzak-Weglarz, E. Docampo, L. E. Duncan, K. Egberts, S. Ehrlich, G. Escaramis, T. Esko, X. Estivill, A. Farmer, A. Favaro, F. Fernandez-Aranda, M. M. Fichter, K. Fischer, M. Focker, L. Foretova, A. J. Forstner, M. Forzan, C. S. Franklin, S. Gallinger, I. Giegling, J. Giuranna, F. Gonidakis, P. Gorwood, M. G. Mayora, S. Guillaume, Y. Guo, H. Hakonarson, K. Hatzikotoulas, J. Hauser, J. Hebebrand, S. G. Helder, S. Herms, B. Herpertz-Dahlmann, W. Herzog, L. M. Huckins, J. I. Hudson, H. Imgart, H. Inoko, V. Janout, S. Jimenez-Murcia, A. Julia,

G. Kalsi, D. Kaminska, J. Kaprio, L. Karhunen, A. Karwautz, M. J. H. Kas, J. L. Kennedy, A. Keski-Rahkonen, K. Kiezebrink, Y. R. Kim, L. Klareskog, K. L. Klump, G. P. S. Knudsen, M. C. La Via, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat Genet*, 51(8):1207–1214, 2019.

[216] M. R. Watson, C. T. Ward, A. Prabhakar, B. Fiza, and V. Moll. Successful use of octreotide therapy for refractory levofloxacin-induced hypoglycemia: A case report and literature review. *Case Rep Crit Care*, 2019:3560608, 2019.

[217] 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.

[218] H Wickham. *ggplot2: Elegant Graphics for Data Analysis*. 2009.

[219] 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.

[220] 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.

[221] Z. Yilmaz, M. Halvorsen, J. Bryois, D. Yu, L. M. Thornton, S. Zerwas, N. Micali, R. Moessner, C. L. Burton, G. Zai, L. Erdman, M. J. Kas, P. D. Arnold, L. K. Davis, J. A. Knowles, G. Breen, J. M. Scharf, G. Nestadt, C. A. Mathews, C. M. Bulik, M. Mattheisen, J. J. Crowley, and Tourette Syndrome Obsessive-Compulsive Disorder Working Group of the Psychiatric Genomics Consortium Eating Disorders Working Group of the Psychiatric Genomics Consortium. Examination of the shared genetic basis of anorexia nervosa and obsessive-compulsive disorder. *Mol Psychiatry*, 2018.

[222] 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.

[223] N. Y. Yu, B. M. Hallstrom, L. Fagerberg, F. Ponten, H. Kawaji, P. Carninci, A. R. Forrest, Consortium Fantom, Y. Hayashizaki, M. Uhlen, and C. O. Daub. Complementing tissue characterization by integrating transcriptome profiling from the human protein atlas and from the fantom5 consortium. *Nucleic Acids Res*, 43(14):6787–98, 2015.

[224] F. Yue, Y. Cheng, A. Breschi, J. Vierstra, W. Wu, T. Ryba, R. Sandstrom, Z. Ma, C. Davis, B. D. Pope, Y. Shen, D. D. Pervouchine, S. Djebali, R. E. Thurman, R. Kaul, E. Rynes, A. Kirilusha, G. K. Marinov, B. A. Williams, D. Trout, H. Amrhein, K. Fisher-Aylor, I. Antoshechkin, G. DeSalvo, L. H. See, M. Fastuca, J. Drenkow, C. Zaleski, A. Dobin, P. Prieto, J. Lagarde, G. Bussotti, A. Tanzer, O. Denas, K. Li, M. A. Bender, M. Zhang, R. Byron, M. T. Groudine, D. McCleary, L. Pham, Z. Ye,

S. Kuan, L. Edsall, Y. C. Wu, M. D. Rasmussen, M. S. Bansal, M. Kellis, C. A. Keller, C. S. Morrissey, T. Mishra, D. Jain, N. Dogan, R. S. Harris, P. Cayting, T. Kawli, A. P. Boyle, G. Euskirchen, A. Kundaje, S. Lin, Y. Lin, C. Jansen, V. S. Malladi, M. S. Cline, D. T. Erickson, V. M. Kirkup, K. Learned, C. A. Sloan, K. R. Rosenbloom, B. Lacerda de Sousa, K. Beal, M. Pignatelli, P. Flicek, J. Lian, T. Kahveci, D. Lee, W. J. Kent, M. Ramalho Santos, J. Herrero, C. Notredame, A. Johnson, S. Vong, K. Lee, D. Bates, F. Neri, M. Diegel, T. Canfield, P. J. Sabo, M. S. Wilken, T. A. Reh, E. Giste, A. Shafer, T. Kutyavin, E. Haugen, D. Dunn, A. P. Reynolds, S. Neph, R. Humbert, R. S. Hansen, M. De Bruijn, et al. A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527):355–64, 2014.

[225] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4:Article17, 2005.

[226] 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.

[227] Aaron M. Zorn and James M. Wells. Vertebrate endoderm development and organ formation. *Annual Review of Cell and Developmental Biology*, 25:221–251, 2009.