THE UNIVERSITY OF CHICAGO

INVESTIGATING THE DYNAMIC REGULATION OF GENE EXPRESSION DURING CELLULAR DIFFERENTIATION

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

KATHERINE LOUISE RHODES

CHICAGO, ILLINOIS

MARCH 2021

Table of Contents

Tables S2-1, S2-5, and S2-8 are available as supplementary files online. The List of Tables gives the page number for each table's caption.

List of Figures

List of Tables

Acknowledgements

Earning a PhD is not an individual achievement, nor is it a symbol of superior intellect, hard work, or even ambition, as I would have thought 6 years ago. A PhD is actually the manifestation of nurtured curiosity and perseverance, which is only made possible by a community of mentors, role models, friends, and family. So, consider this the most important section of this thesis.

First, I'd like to thank my advisor, Dr. Yoav Gilad. Yoav has had an enormous impact on my scientific thinking and has shaped the way I approach experiments, the way I ask scientific questions, and the way evaluate assumptions and preconceptions. Yoav has also managed to build a positive and supportive lab environment, with the right balance of structure and freedom – an environment that enabled me to grow as both a person and as a scientist. I'd also like to thank my committee members Dr. Anindita Basu, Dr. Matthew Stephens, and Dr. Marcelo Nobrega whose guidance and perspective have improved the quality of my work. And I'd like to thank the administrators Sue Levison and Tamiko Charley for their work and support; they makes these departments, committees, and programs run.

I also need to thank my amazing labmates and collaborators. No one has shaped my grad school experience more than Reem Elorbany—my lab twin, the other head of my doduo, the brain to my pinky. We worked side by side from our very first day in lab and shared both our failures and our successes. Thanks also to the third member of 'the cohort' Briana Mittleman whose passion and motivation are truly infectious. Briana and Reem supported me through the day-to-day struggles of grad school both in and out of lab and I'm a better person for knowing them. Thanks to Kenneth Barr for sharing his expertise and diligence as we worked to get the

embryoid body system off the ground. Thanks to Lauren Blake and Ittai Eres for being my peer mentors. Lauren and Ittai have constantly provided generous advice and introduced me to opportunities all over the university. Thank you to all the other graduate students, undergraduate student, postdocs, and technicians in the Gilad Lab including Bryan Pavlovic, John Blischak, Anthony Hung, Deji Adegunsoye, Wenhe Lin, Erik McIntire, Shane Warland, Po-Yuan Tung, Joyce Hsiao, Michelle Ward, Genevieve Housman, Ben Fair, Ben Umans, Claudia Cuevas, Emilie Briscoe, Jonathan Burnett. Thanks also to Sebastian Pott, Natalia Gonzales, Abhishek Sarkar, and Peter Carbonetto. And thank you to my collaborators in Dr. Alexis Battle's lab: Ben Strober, Josh Popp, Karl Tayeb, and Nirmal Krishnan.

I am also grateful to my friends in my Human Genetics cohort: Liana Hernandez, Olivia Gray, Kevin Magnaye, Joe Marcus, Charles Washington, Arjun Biddanda, Manny Vasquez, Brandon Mapes, and Nan Xiao. I was in awe of all of them when we first began, and I continue to be awed by their accomplishments. It has been a privilege to learn with them and from them.

Thank you to Simone Rauch, my friend, doppelganger, and forever-roommate (even when we live on opposite sides of the Atlantic). Simone was the first person I met when I visited UChicago on my interview weekend—we shared a fateful UGo shuttle from O'Hare to Hyde Park—and I'm so grateful to have had her friendship and support from the very beginning.

Thank you to my dungeons and dragons parties and to my improv groups. With these friends, I've been thousands of people, lived countless adventures, told myriad stories, and just really had so much fun. In particular, I want to thank Charles Washington for introducing me to D+D (well, pathfinder) and continuing to put-up with my chaotic characters. Thank you to Allison Haungs, Zaina Zayyad, Reem Elorbany, Joe Marcus, Arjun Biddanda, Liana Hernandez,

John Park, Talia Berger, Laurel Ann Gonsecki, Louisa Gonsecki, and Quincey Smail for being huge nerds. And thank you to improv freaks Victoria Bujny, Amy Tseui, Paul Brandt, William Starkoff, Benjamin Yoder, Frederick Nitsch, Mark Bon, and Jonathan Pfendler. They're a bunch of weirdos but they've got my back. I also want to thank a few of my improv and sketch teachers, Andel Sudik, John Hildreth, Dina Facklis, Farrell Walsh, and Jorin Garguillo, for encouraging me to be silly, to listen, and to trust my instincts.

Last but not least, I want to thank my family. I couldn't –and wouldn't--have done this without their love and support. My parents, Diane Dietzen and Anthony Rhodes, have instilled in me a love of science since I was small. My dad likes to tell this story about a time when we went to the Academy of Natural Sciences in Philadelphia when I was, like, 4 or 5 and I was riding around on his shoulders pointing to each dinosaur skeleton and loudly exclaiming the exact genus and species while other museum-goers looked on in awe. I have to imagine that my dinosaur expertise is *slightly* exaggerated in that story, but I can also draw a straight line from that story to where I am today. Thank you to my brother, John Rhodes. No one else can quote my favorite TV shows line by line with me. John has also shown me the true meaning of strength and resilience. Thank you to my grandmothers Mary Rhodes and Sandra Dietzen who have taught me to be a kind, ambitious, and outspoken woman. And thank you to aunts, uncles, and cousins—I've learned so much from each of you. I love you all and I owe this achievement to you.

Abstract

Gene regulation is context-specific and dynamic, changing between cell types and states, and during processes like differentiation and development. To more deeply understand the genetic architecture of human traits and diseases, it is necessary to characterize genetic effects on gene regulation in the relevant cell types and in dynamic contexts. In this thesis, I addressed this goal using *in vitro* differentiations of human induced pluripotent stem cells (iPSCs) to study dynamic regulation of gene expression. In Chapter 2, I differentiated a panel of iPSCs to cardiomyocytes and collected timecourse RNA-seq data at regular intervals throughout the differentiation. I used this data to identify dynamic eQTLs, or genetic loci whose effect on gene expression changes through time. I found that dynamic eQTLs acting during cardiomyocyte differentiation are associated with heart phenotypes and cardiovascular disease risk. In Chapter 3, I characterized patterns of gene expression in embryoid bodies (EBs), a differentiation system that enables the simultaneous study of a multitude of cell types and differentiation trajectories. I found that EBs are composed of functionally and temporally diverse cells, including cells with similar transcriptomes to primary fetal cell types. This work provides a foundation of knowledge of the EB system that will enable future studies of dynamic eQTLs. In chapter 4, I generated EBs from human and chimpanzee lines to compare gene expression between species across many cell types, including early developmental cell types which have previously been inaccessible. I identified patterns of conserved and diverged gene expression between species. Overall, these studies leverage iPSC differentiations to study dynamic gene regulation and provide new insight into the genetics of human development, complex traits, and disease.

Chapter I: Introduction

A major goal of human genetics and genomics is to understand the genetic architecture of human traits and disease. Over the past 20 years, genome wide association studies (GWAS) have identified thousands of trait-associated loci (1, 2). In cases where these loci lie in protein-coding regions, it is relatively easy to predict the effect of a single nucleotide polymorphism (SNP) on protein structure and biochemistry (3, 4). However, the majority of variants identified in GWAS are located in non-coding regions of the genome and are hypothesized to impose phenotypic consequences by affecting gene regulation (5–7). Unlike protein-coding regions of the genome, it is difficult to predict the impact of genetic variation in non-coding regions based solely on the DNA sequence. In fact, it is challenging to even predict which genes are regulated by any given noncoding SNP.

One approach to characterizing the regulatory effects of non-coding variants is to perform quantitative trait locus, or QTL, analysis. A QTL is a location in the genome where genetic variation is associated with variation in a quantitative trait. Any heritable, continuous phenotype that can be accurately measured can be used in QTL analysis, including regulatory phenotypes such as protein levels, DNA methylation, chromatin accessibility, and gene expression (8–11). To characterize the regulatory functions of non-coding GWAS variants, one can ask whether a certain locus identified in a GWAS is also a QTL for a regulatory phenotype (12). For example, suppose we identify a certain non-coding SNP as being strongly associated with risk for myocardial infarction in a GWAS. And, we identify the same SNP as a chromatin accessibility QTL in heart tissue. We may conclude that this SNP influences long-term risk of heart attack by functionally altering the chromatin landscape in heart tissue.

QTLs associated with gene expression (eQTLs), in particular, have been widely studied and efforts to map eQTLs have included large-scale, coordinated analyses by the Genotype-Tissue Expression (GTEx) Consortium(13). GTEx has collected RNA sequencing and genotype data from 54 adult tissues and cell lines representing over 800 individuals; using this enormous sample size, GTEx has had the statistical power to identify over 4.2 million eQTLs (14). Yet, only 11% of disease heritability can be attributed to GTEx cis eQTLs, and only 43% of disease-associated variants identified by GWAS are also classified as eQTLs (15). Altogether, this illustrates that a significant fraction of genetic disease risk can be attributed to genetic effects on gene regulation. But, the majority of GWAS variants still remain unexplained, even using the comprehensive GTEx map of eQTLs (16).

Genomic studies of dynamic gene regulation

Gene regulation is dynamic, changing between cell types, between environmental conditions, and throughout processes like differentiation and development (*17*, *18*). It is possible, then, that the reason there are so many unexplained non-coding disease-associated variants is because we have not been assaying gene expression in the full array of relevant contexts. GTEx, for example, has mapped eQTL across diverse tissue types, but these tissues are collected post-mortem, mostly from adults. GTEx, then, will miss genetic effects on gene regulation that only occur in earlier life stages. To continue the effort to understand the regulatory behavior of all disease associated genetic variants, it will be necessary to identify eQTLs in previously unstudied tissues and cell types, including early developmental cell types.

Another limitation of GTEx is that, because their data set only captures mature tissues, their analyses are restricted to identifying standard, static, eQTLs. In other words, they assess a tissue's transcriptome in only one state, rather than capturing the response to an environmental

exposure or change through time. Static eQTLs are typically shared across all tissues, and probably do not contribute to disease in a context-specific matter (*13, 17*). Recently, dynamic eQTLs – sometimes referred to as response eQTLs—have emerged as an alternative and complementary approach to static eQTL mapping. Dynamic eQTLs are loci associated with the change in gene expression level through time during a process of interest such as treatment response or differentiation. Due to the fact that dynamic eQTLs, by definition, have context-specific effects, they may be more likely to have a significant deleterious effects compared to static eQTLs with ubiquitous regulatory effect across all tissues (*19*). Indeed, recent studies of dynamic eQTL have successfully revealed the regulatory mechanisms of disease-associated variants (*20–23*). In order to more fully understand complex phenotypes and disease mechanisms, we must characterize eQTLs not only in diverse, disease-relevant cell types but also within dynamic temporal and environmental contexts.

iPSCs and iPSC-derived cell types as a model system

In vitro studies of human development can be accomplished using induced pluripotent stem cells, or iPSCs, as a model system. iPSCs are cells that have been reprogrammed from adult somatic cells to return them to a pluripotent state (*24*). This reprogramming can be a achieved via transient overexpression of a few transcription factors, and can be done using a variety of somatic cell types as a starting point, including easily accessible cell types like skin fibroblasts and transformed B cells from whole blood (*25–27*). Because these cell types can be collected non-invasively, it is possible to ethically generate human iPSCs from practically any individual who provides their informed consent.

iPSCs are an advantageous *in vitro* model system for several key reasons. First, they are capable of self-renewing indefinitely and can be cryopreserved, making it possible to grow and

bank large volumes of cells for future experimental use. Second, iPSCs can be directedly differentiated into many cells types including cardiomyocytes, hepatocytes, and neurons using well-established and reproducible protocols (*28–30*). Cardiomyocytes (CMs) are perhaps the most experimentally tractable iPSC-derived tissue type in that the differentiation process is relatively quick, producing beating CMs in about a week, and protocols for CM differentiation are well established and widely used (*28, 31*). Importantly, both the iPSCs and iPSC-derived terminal cell types faithfully recapitulate most gene expression patterns found in primary tissue samples (*32*). Furthermore, patterns of inter-individual variation in gene regulation have been shown to be maintained when somatic cells are reprogrammed into iPSCs and when iPSCs are differentiated into terminal cell type (*27*); hence, QTL studies performed in these cell types can yield biologically meaningful insight into inter-individual differences in gene regulation.

In vitro differentiation of iPSCs also enables analysis of immature, developmental cell types and developmental regulatory dynamics. Genomic studies of human development *in vivo* have historically been difficult or impossible due to the inaccessibility of human fetal tissues. With iPSCs, not only can we assay developmental cell types, but we can capture cells throughout the differentiation process, essentially using differentiation time as a treatment condition to identify transient and fleeting regulatory events. Additionally, studies that incorporate large sample sizes and high temporal resolution of iPSC differentiation can identify dynamic eQTLs.

There are, however, several logistical barriers for studies of directed differentiations, or protocols that induce iPSCs to develop into a relatively homogeneous population of a certain cell type. To date, directed differentiation protocols have been developed for a variety of cell types (*30, 33–40*). Unfortunately, the number of cell types that can be generated using these protocols is extremely limited compared to the diversity of cell types that

make up the human body. Moreover, the efficacy of each protocol varies and the resulting variation in cell type composition can impede analysis and decrease power to detect QTL(*41*). Directed differentiations also tend to be expensive, inefficient, and laborious. These challenges restrict the breadth of differentiations and developmental trajectories in which dynamic QTLs can be identified in a practical manner. iPSC differentiations that result in heterogeneous cell types, such as organoid differentiations of embryoid body differentiations, can overcome many of these barriers thereby enabling the efficient analysis of many differentiation trajectories in a single study(*37*, *42–45*).

The future of embryoid bodies in genomic studies

Embryoid bodies (EBs) are three dimensional aggregates of iPSCs spontaneously differentiating to cell types of all three germ layers (*46*). EB formation has been used to test stem cells for pluripotency for decades. Until recently, the complexity of EBs has precluded their use as genomic models. EBs do not produce pure cell populations that can be meaningfully analyzed with bulk RNA-seq data; instead, EBs are composed of highly heterogeneous cell types representing diverse functionalities across many stages of differentiation. With the advent of single-cell RNA sequencing, it is now possible to computationally deconvolve EBs into their component cell types. This strategy provides a way to map eQTLs in a multitude of cell types arising from the same genotype, including developmental cell types that would otherwise be inaccessible either due to the ethical limitation of obtaining primary fetal tissue or due to lack of a known directed differentiation protocol. EBs are also an incredibly efficient and practical model system: a multitude of diverse cell types can be generated in a single dish, enabling strict control of environmental and technical factors. Complex EBs can be generated in only a few weeks, limiting the amount of time, money, and labor required for experiments. With these

advantages, EBs have the potential to be a powerful model system for studies of human developmental gene regulation. As with any model system, foundational work is required to understand the biological and technical variation contributing the resulting datasets. The study detailed in chapter 3 of this thesis builds this critical foundation, and will support the design and implementation of high powered studies of gene expression variation in the future.

EBs can also enable studies of gene regulation across diverse cell types in non-human primates and other non-model organisms. For many non-human species, collection of tissues, even post-mortem, is legally prohibited (*47*, *48*). Thus, studies of gene regulation have been limited to tissues collected before legal restrictions were imposed. For species where non-invasive tissue collection has been legally achieved, iPSCs can be reprogrammed. In turn, EBs can be generated and analyzed with single cell RNA-sequencing (scRNA-seq), facilitating the study of diverse tissues without unethical, invasive tissue collection. To date, iPSCs have been generated from an assortment of species, including, but not limited to, nonhuman primates (including chimpanzees and rhesus macaques), horses, dogs, pigs, cows, naked mole rats, and even endangered species (including mandrills and white rhinos)(*49*, *50*). The diverse cell types made available by EB differentiation can deepen our understanding of gene regulation in these species, and can be used in comparative analyses to understand conserved and human-specific patterns of gene regulation.

Comparative approaches to understanding gene regulation

Comparative studies of functional genomics between humans and nonhuman primates have significantly contributed to our understanding of human gene regulation(*19*, *51–54*). Over the last decade, dozens of comparative genomic studies have characterized differences in regulatory phenotypes between humans and primates in an array of tissues. Together, these

studies have constructed large catalogs of genomic data sets measuring gene expression levels, DNA methylation levels, chromatin modifications and accessibility, and even protein expression levels which have yielded novel insights into the nature of gene regulation(*51, 55–60*).

Notably, sample sizes in comparative studies tend to be small due to the difficulty in obtaining primary tissues from nonhuman primates. These small sample sizes it difficult to perform QTL mapping. But, it is still possible to perform statistically powerful differential expression (or differential methylation, differential chromatin accessibility, etc.) analysis due to the relatively large degree of genetic variation and subsequent large effect sizes between species. Indeed, even with small sample sizes, comparative studies of gene regulation have yielded meaningful results and novel insights into human gene regulation. For example, Eres et al. performed a study using hi-C to measure 3D chromatin conformation in a small sample of 4 human and 4 chimp iPSCs. Hi-C is a complex, laborious, and expensive protocol, making it technically and financially impossible to include many individuals (*61*). Due to the large effect sizes between species, Eres et al. were nevertheless able to demonstrate that differences in contact frequencies at certain genomic loci between species likely underlie species-specific patterns of gene expression (*61*). More broadly, this revealed the novel insight that variation in chromatin conformation mediates differences in gene expression. Much larger sample sizes would be needed to draw the same conclusions in a sample of only human cells.

Comparative genomic studies can also contextualize gene regulatory mechanisms in an evolutionary framework to identify conserved and derived patterns of human gene regulation. By identifying conserved patterns of gene regulation we can reveal cellular processes that are critical for organismal function and have thus been maintained through evolution via purifying selection (*19*). Conserved processes often include developmental pathways and pathways

involved in cellular maintenance. Of course, it is also exciting to learn about derived patterns of gene regulation that distinguish humans and our nearest ancestors. Most of the genetic differences between humans and chimpanzees, for example, lie in noncoding regions of the genome and are thought to contribute to phenotypic differences by affecting gene regulation (*62*, *63*). When differential gene regulation is observed, it can be tempting to draw vast conclusions about the nature of being human and to construct evolutionary stories as to why we got to be this way. We should, however, be cautious in our interpretation of observed differences in gene regulation between species. Because sample sizes are often small, observed differences could be due to sample bias. Moreover, in cases where primary tissue was collected under non-optimal conditions, observed differences could be due to uncontrolled technical or environmental factors(*54*). Additionally, there are many steps in the regulatory cascade between DNA sequence variation and large scale phenotypic variation. So, for example, inter-specific differences at gene expression level may be buffered at the level of protein expression and, consequently, may not have any effect on morphological differences between species (*64*). While comparative studies are a useful tool for understanding gene regulation, interpretation of differences between species should be done thoughtfully and cautiously.

Dissertation Overview

In this thesis I investigated the dynamics of gene regulation during cellular differentiation using iPSCs as a model system. Specifically, I first used bulk RNA-sequencing in a high-resolution time course of iPSC-derived cardiomyocyte differentiation to identify dynamic eQTLs -- locations in the genome where natural variation between individuals is associated with differences in gene expression which are modulated by differentiation time. I then explored the relationship of these dynamic eQTLs with variants associated with cardiovascular traits and

disease to identify dynamic eQTLs with transient effects during development that may have a long-term impact on human health. Next, I differentiated embryoid bodies from a small sample of human and chimp iPSC lines. Using just the human EBs, I quantified the contribution of biological and technical factors to overall patterns of gene expression, identified the functionally and temporally diverse cell types present, and explored inter-individual patterns of dynamic gene expression in diverse developmental trajectories. Overall, these analyses establish human embryoid bodies as a powerful model system for genomic studies. Last, I performed a comparative analysis of human and chimp EBs. I characterized differential expression across all EB cell types and identified conserved patterns of gene regulation between species. This work also produced a reference set of chimpanzee cell types which cannot be studied *in vivo*.

Chapter II: Dynamic genetic regulation of gene expression during cardiomyocyte differentiation

Note:

The following section (*Chapter II*) is reproduced verbatim, with the exception of chapter title, figure numbering, and reference labeling, from my co-first authored reference "Dynamic genetic regulation of gene expression during cellular differentiation" (Strober et al. 2019). This project was performed in collaboration with Benjamin Strober and Katherine Rhodes, and published in *Science* on June 28, 2019.[1]

Authors:

B. J. Strober*, R. Elorbany*, K. Rhodes*, N. Krishnan, K. Tayeb, A. Battle[+], and Y. Gilad[+]

*These authors contributed equally to this work.

Abstract

Genetic regulation of gene expression is dynamic, as transcription can change during cell differentiation and across cell types. We mapped expression quantitative trait loci (eQTLs) throughout differentiation to elucidate the dynamics of genetic effects on cell type–specific gene expression. We generated time-series RNA sequencing data, capturing 16 time points during the

[1] Strober, B. J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., & Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. Science, 364(6447), 1287–1290. Reprinted with permission from AAAS.

differentiation of induced pluripotent stem cells to cardiomyocytes, in 19 human cell lines. We identified hundreds of dynamic eQTLs that change over time, with enrichment in enhancers of relevant cell types. We also found nonlinear dynamic eQTLs, which affect only intermediate stages of differentiation and cannot be found by using data from mature tissues. These fleeting genetic associations with gene regulation may explain some of the components of complex traits and disease. We highlight one example of a nonlinear eQTL that is associated with body mass index.

Full text

Genetic variants that alter gene regulation play an essential role in the genetics of human disease and other complex phenotypes (*65*, *66*). Large studies have identified thousands of genetic loci associated with complex diseases, most of which are in noncoding regions of the genome and therefore are putatively involved in gene regulation (*66*). Expression quantitative trait locus (eQTL) analysis has shown that many disease-associated loci influence the regulation of nearby genes (*67*, *68*) but a substantial fraction of disease-associated loci still remain unexplained (*69*, *70*).

Much effort has been dedicated to mapping and identifying eQTLs across tissues and cell types, as the regulatory impact of disease-associated loci may be most evident in cell types relevant to each disease. Regulatory genetic effects can also be time point specific or environment dependent (*71*, *72*) and may influence temporal programs of gene regulation. Yet almost all studies of the genetics of gene regulation, including the multitissue Genotype-Tissue Expression (GTEx) project (*71*), involve data collected at a single time point, usually from adult individuals. Dynamic gene expression data can add another dimension to eQTL analysis,

allowing identification of genetic variants with transient effects that may not have been found in analysis of static data.

We took advantage of a panel of induced pluripotent stem cell (iPSC) lines from 19 individuals to investigate high-resolution temporal genetic effects on gene regulation over time during cardiomyocyte differentiation. Specifically, we collected gene expression data throughout the differentiation from iPSCs to cardiomyocytes in 19 well-characterized human Yoruba HapMap cell lines (*32*). For each cell line, RNA was extracted and sequenced every 24 hours for 16 days to capture the entire differentiation process; in total, we sequenced 297 RNA samples (figs. S2-1 and S2-2). Combined with available whole-genome sequences and genotype data for each cell line, these data provide a resource with which to investigate how gene expression and genetic regulation change throughout cardiomyocyte differentiation with high temporal resolution.

During iPSC culturing, differentiation, RNA extraction, and processing for sequencing, we recorded extensive metadata on each sample (table S2-1). Quality controls and filtering yielded 16,319 genes for downstream analysis (*Materials and Methods*). After standardization and normalization of the RNA sequencing (RNA-seq) data (*Materials and Methods*), we evaluated the contribution of potential confounders to overall variation in our data, confirming that our study design was effective (fig. S2-3). We also used replicates from an independent differentiation to confirm that the gene expression patterns we observed in our iPSCs and iPSC-derived cardiomyocytes are robust with respect to variance that may be associated with the differentiation procedure (fig. S2-4) (*32*) (*Materials and Methods*).

We evaluated the efficiency of our differentiation by fluorescence-activated cell sorting (table S2) and by considering the time-course expression of known cell type–specific marker genes (*25*, *73*) (fig. S2-5). As expected (*73*), cardiomyocyte purity and the expression of lineage marker genes are variable across our samples. This variability between cell lines was observed

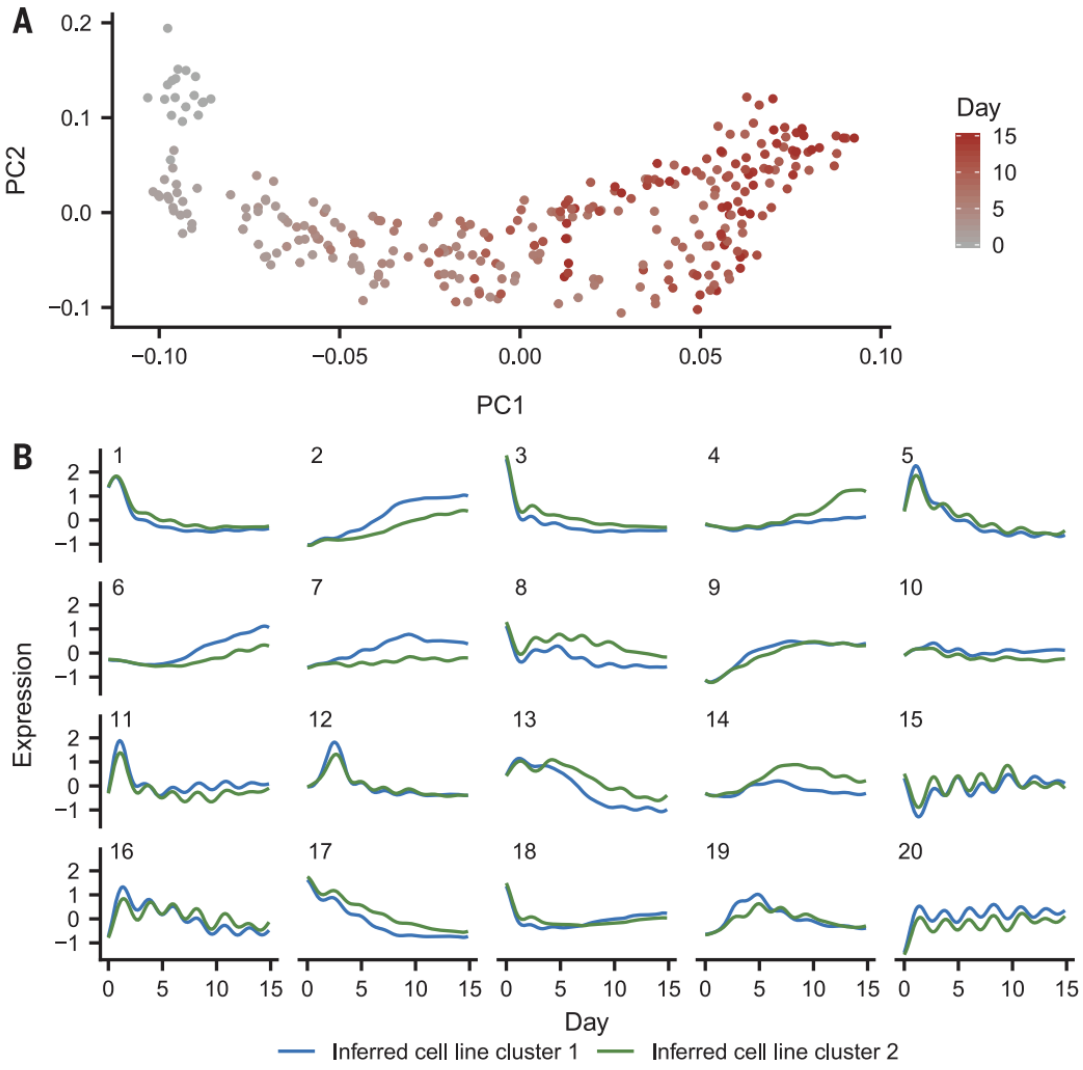


Fig. 2-1. Gene expression trends throughout cardiomyocyte differentiation. (A) The first two gene expression principal component (PC) loadings for all 297 RNA-seq samples across cell lines, where each sample is colored according to the day of collection. (B) Predicted cell line cluster expression trajectories for 20 gene clusters according to split-GPM. Many gene clusters (8, 11, 15, 16, and 20) exhibit periodic expression trajectories that correspond with cell culture media changes.

across the entire time course, although the effect of differentiation time is the primary source of variation in the data (Fig. 2-1A and figs. S2-3 and S2-6).

We characterized global patterns of gene expression across time by applying split-GPM, an unsupervised probabilistic model that infers time-course trajectories of gene expression using Gaussian processes, while simultaneously performing clustering of genes and cell lines (*Materials and Methods*). Using this approach, we identified two clusters of cell lines that displayed broad differences in the expression patterns of multiple clusters of genes; in each gene cluster, genes exhibit shared expression changes over time. The assignment of cell lines to clusters is robust with respect to the parameters we tested, such as the number of inferred gene clusters (fig. S2-7).

The two cell line clusters we identified differ in the efficiency of cardiomyocyte differentiation. Cell lines in the first (larger) cluster display greater troponin expression levels in the final six time points of differentiation ($P = 0.014$, Wilcoxon rank-sum test). The expression of a group of genes enriched for myogenesis also increases by a greater magnitude over time in cell lines in the first cluster (Bonferroni $P = 9.29 \times 10{-}14$) (gene cluster 2 in Fig. 2-1B) (*74*). Cell lines in the second, smaller cluster show high expression of genes related to KRAS activation (Bonferroni $P = 0.005$; gene cluster 4 in Fig. 2-1B), which is associated with increased self-renewal of undifferentiated iPSCs and decreased neuronal differentiation propensity (*75*). Other gene clusters illuminate broad changes in gene expression over time, such as a transient rise in *MYC* and *E2F* target genes in the early days of differentiation (gene cluster 13 in Fig. 1B; table S2-3). Together, this analysis documents patterns of gene expression trajectories over time and captures differences among our cell lines that are not obvious from the individual time point data alone.

Next, we evaluated the impact of genetic variation on gene regulation in our system. We used WASP software (*76*) to identify cis-eQTLs in the data from each time point independently (*Materials and Methods*). To control for latent confounders in the independent analysis of data from each time point, we included the first three expression PCs using data from samples of the corresponding time point as covariates (figs. S2-8 and S2-9, A and B). At an empirical false discovery rate (eFDR) of 5%, we identified a median of 111 genes (range: 71 to 231) with at least one eQTL in each time point (figs. S2-9C and S2-10). As expected, the eQTLs we

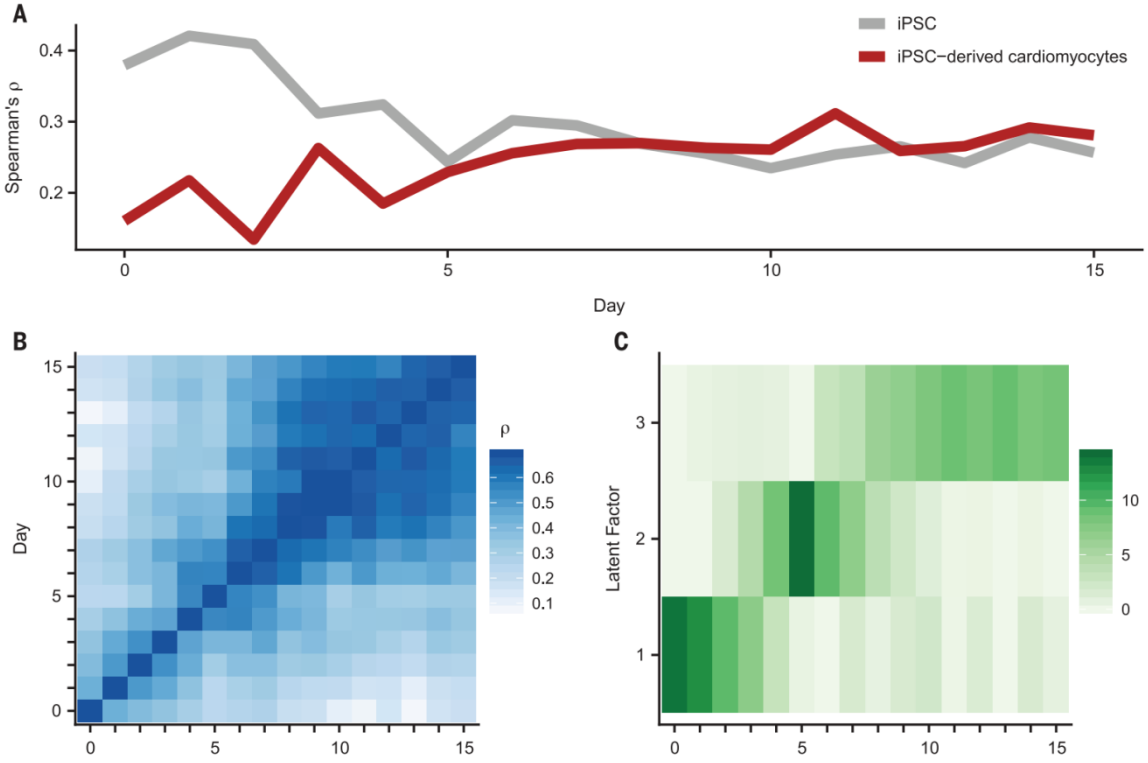


Fig. 2-2. eQTL patterns during cardiomyocyte differentiation. We limit this analysis to genes with at least one significant eQTL (WASP combined haplotype test; eFDR ≤ 0.05) across time points. If a gene has more than one significant eQTL, we select a single variant for that gene with the smallest geometric mean P value across all 16 time points. (A) Spearman correlation of P values between eQTLs from each day (x axis) and existing iPSC (gray) and iPSC-derived cardiomyocyte (red) eQTLs. (B) Spearman correlation of eQTL P values for each pair of days. (C) Factors identified via sparse matrix factorization of eQTL-log10 P values using three latent factors and an L1 penalty of 0.5.

identified early in the time course replicated in data from iPSCs, whereas eQTLs from later time points were better supported by data from iPSC-derived cardiomyocytes (both $P < 0.001$, linear regression) (Fig. 2-2A) (*32*).

We computed the correlation of the significant eQTL summary statistics for each pair of time points (Fig. 2-2B). We observed that correlation between eQTL summary statistics increases as the distance between time points decreases ($P \leq 2 \times 10{-}16$, linear regression). Although this observation is intuitive, it indicates that the dynamic impact of genetic variation on gene regulation in our data is not random and is related to the temporal process of cardiomyocyte differentiation.

To more formally quantify the temporal structure of genetic regulation throughout differentiation, we performed sparse non-negative matrix factorization on the matrix of significant eQTL summary statistics from all time points (*Materials and Methods*). The learned factors capture genetic signal that is largely specific to a subset of differentiation time (Fig. 2-2C), a pattern that is robust with respect to the number of latent factors or sparse prior choice (fig. S2-11).

Our analysis indicates that temporal structure dominates the patterns of genetic association with gene expression in our data. However, the observation that most significant nondynamic eQTLs can be identified in only a few time points (median of 2) (fig. S2-12) is most likely explained by incomplete power to identify eQTLs in each time point independently. To robustly identify dynamic eQTLs whose effect varies significantly over time, leveraging power across all time points (Fig. 2-3A), we used a Gaussian linear model applied jointly to data from the entire experiment. Specifically, we quantified the effect of interactions between genotype and

differentiation time on gene expression, controlling for linear effects of both differentiation time and genotype. In addition, we accounted for the systematic differences in differentiation trajectories identified between cell lines (Fig. 2-1B, figs. S2-13 to S2-16, and table S2-4) (*Materials and Methods*), which would otherwise lead to false positives in our analysis. Using this approach, we identified 550 genes with a significant dynamic eQTL (eFDR ≤ 0.05) (figs. S2-17 to S2-20 and table S2-5).

We classified the 550 dynamic eQTL as early (eQTL effect size decreasing over time), late (eQTL effect size increasing over time), or switch (eQTL effect size exhibiting different directions of effect over time) (fig. S2-21) (*Materials and Methods*). We found that the early dynamic eQTLs are enriched for chromHMM enhancer elements annotated in iPSC Roadmap Epigenomics cell types but not in heart-related cell types (*77*, *78*). In turn, late dynamic eQTLs are enriched for chromHMM enhancer elements annotated in heart-related Roadmap Epigenomics cell types but not in iPSCs (Fig. 2-3B and fig. S2-22). These observations indicate that dynamic eQTL mapping can capture temporal changes in cellular gene regulation reflecting changes in regulatory element activity as the cell cultures differentiate.

The observation that we are able to capture the function of cell type–specific regulatory elements prompted us to consider dynamic eQTLs in other contexts. We found that dynamic eQTLs are enriched for genes with roles in myogenesis (Bonferroni P = 0.0019, Fisher's exact) (table S2-6) (*74*) and also show significant enrichment for genes related to dilated cardiomyopathy (P = 0.001, Fisher's exact) (table S2-7) (*79*) (*Materials and Methods*). Two significant dynamic eQTLs in particular, rs7633988 and rs6599234 (in strong linkage disequilibrium; coefficient of determination, R 2 = 0.93), are genome-wide association study variants for QRS duration and QT interval, respectively (fig. S2-23) (*80*, *81*). Both variants show

an association with the expression levels of *SCN5A*, which is involved in the creation of sodium channels and is in the dilated cardiomyopathy gene set (*82*). Another dynamic eQTL, rs11124033, associated with the expression of *FHL2* (Fig. 2-3A), is also associated with dilated cardiomyopathy. This variant lies in a Roadmap Epigenomics chromHMM promoter element annotated in heart-related cell types but not in iPSCs (*77*, *78*). None of these examples were identified as eQTLs in the nondynamic QTL analysis of each time point from our dataset or in the GTEx heart tissue data (*71*).

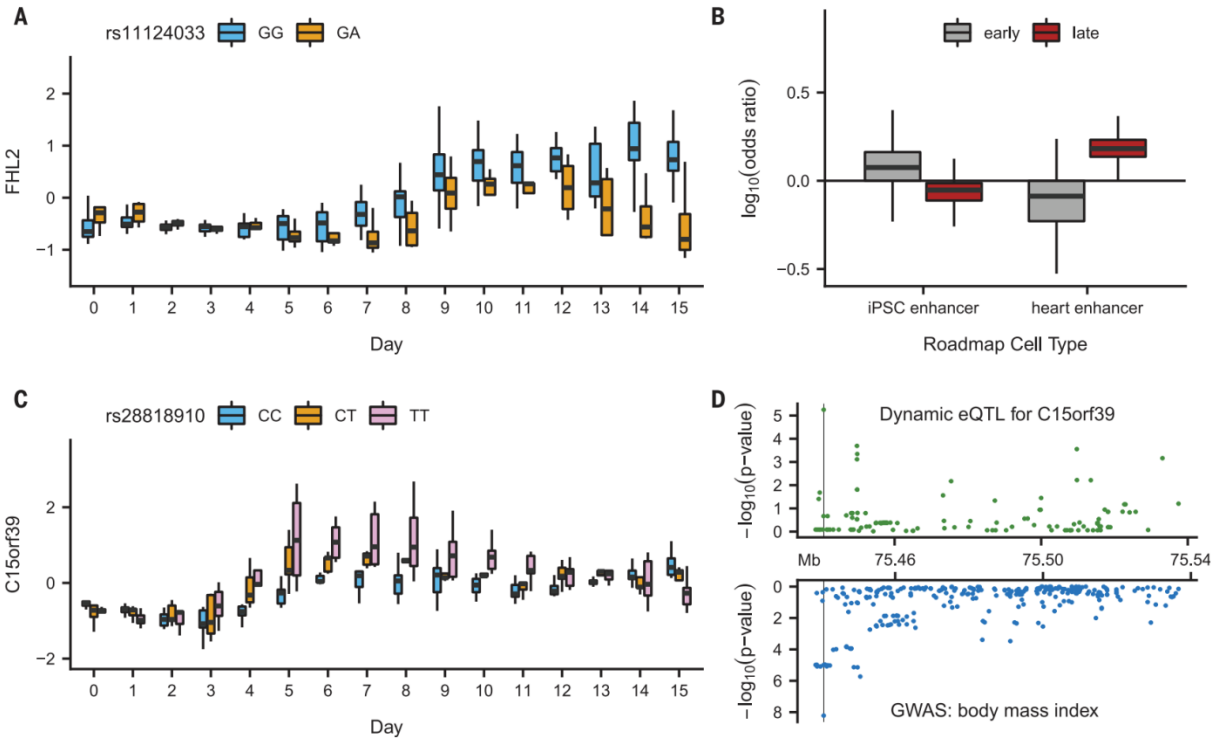


Fig. 2-3. Dynamic eQTLs detect genetic regulatory changes caused by cardiomyocyte differentiation. (A) Linear interaction association between genotype (color) of rs11124033 and time point (x axis) on residual gene expression (cell line effects regressed on expression) of *FHL2* (y axis). (B) Enrichment of dynamic eQTLs in cell type–specific chromHMM enhancer elements relative to 1000 sets of randomly selected matched-background variants. Dynamic eQTLs were classified as either early or late. (C) Nonlinear interaction association between genotype (color) of rs28818910 and time point (x axis) on residual gene expression of *C15orf39* (y axis). (D) Nonlinear interaction association significance of all variants tested within 50 kb of the *C15orf39* transcription start site with expression of *C15orf39* (green) and GWAS significance for BMI of variants in the same window (blue). Vertical line depicts genomic location of the most significant nonlinear dynamic eQTL (rs28818910) for *C15orf39*.

Finally, we sought to identify a wider range of dynamic regulatory patterns, including nonlinear associations, such as when a genetic effect increases in magnitude in the middle of the time course before decreasing or disappearing. To identify nonlinear dynamic eQTLs, we expanded our linear model using a second-order polynomial basis function (*Materials and Methods*). We acknowledge that our study is underpowered to expand to a more general class of nonlinear dynamic eQTLs that do not assume a continuous effect of differentiation time (fig. S2-24) (*Materials and Methods*).

We identified 693 genes with a nonlinear dynamic eQTL (eFDR ≤ 0.05) (figs. S2-17B and S2-19B and table S2-8), 28 of which have their strongest genetic effect in the middle of the differentiation time course (middle dynamic eQTLs) (fig. S2-25) (*Materials and Methods*). Twenty-five of these middle dynamic eQTL genes and their strongest associated variant are not identified as eQTLs in our nondynamic QTL analysis in either iPSCs (day 0) or cardiomyocytes (day 15).

In one example of a nonlinear dynamic eQTL, rs8107849 is associated with the expression of *ZNF606* with a larger magnitude of effect during days 4 through 11 (fig. S2-26). The rs8107849 locus does not lie in iPSC or heart-related chromHMM regulatory regions and was not identified in our analysis as a nondynamic eQTL at any time point. Although *ZNF606* is known to have a role in differentiation of chondrocytes (*83*), it is possible this is a conserved process involved in the differentiation of additional cell types, including cardiomyocytes. Another nonlinear dynamic eQTL reveals an association between rs28818910 and *C15orf39*. The rs28818910 variant is also associated with body mass index (BMI) ($P < 6.07 \times 10{-}9$, reported) (Fig. 2-3, C and D) (*84*) and weakly associated with red blood cell count ($P < 1.48 \times 10{-}6$,

reported) (*85*). This dynamic eQTL and both traits show similar patterns of association across the region (fig. S2-27). The rs28818910 locus is associated with inter-individual differences in gene expression only during intermediate stages of differentiation; it does not lie in annotated regulatory elements of either iPSCs or cardiomyocytes and is not identified as an eQTL in iPSCs, mature cardiomyocytes, or either of the two GTEx heart tissues. Thus, this is an example of a temporary dynamic regulatory effect that may have phenotypic consequences.

Our time-course study design allowed us to identify hundreds of dynamic eQTLs throughout the differentiation of human iPSCs to cardiomyocytes. Dynamic eQTLs, in particular those with nonlinear effects, may often be transient and will not be found in studies that only consider gene expression data from either stem cells or mature tissues and cell types. Many of our dynamic eQTLs lie in regions without known regulatory annotations, as functional studies have focused on static cell types. Thus, these loci may have previously unknown regulatory effects, which could be followed up with further functional validation in relevant intermediate time points. The dynamic genetic effects identified in our study, or in future time-series genomic datasets, will provide a resource for investigating mechanisms underlying disease associations that cannot be characterized based on studies of terminal cell types.

Materials and Methods

Samples. We used induced pluripotent stem cell (iPSC) lines from 19 individuals from the Yoruba HapMap population. The iPSC lines were reprogrammed from LCLs and characterized previously (*32*). All 19 individuals are female and unrelated. We chose to use only female individuals to avoid introducing additional variance that is not of interest in this study.

iPSC Maintenance. Feeder-free iPSC cultures were maintained on Matrigel Growth Factor Reduced Matrix (CB40230, Thermo Fisher Scientific) with Essential 8 Medium (A1517001, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). Cells were grown in an incubator at 37°C, 5% CO2, and atmospheric O2. Cells were passaged to a new dish every 3-5 days using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded with ROCK inhibitor Y-27632 (ab120129, Abcam).

Cardiomyocyte Differentiation. We differentiated iPSCs using a protocol previously optimized for use with the Yoruba HapMap panel (*32*). This protocol implements slight modifications to the cardiomyocyte differentiation protocols from Lian et al. 2013 and Burridge et al. 2014. Feeder-free iPSCs were seeded onto wells of a 6-well plate and grown for 3-5 days prior to differentiation. When most lines were 70%-100% confluent, E8 media was replaced with "heart media" along with 1:100 Matrigel hESC-qualified Matrix (08-774-552, Corning) and 12uM of GSK-3 inhibitor CHIR99021 trihydrochloride (4953, Tocris). "Heart media" is composed of RPMI (15-040-CM, Thermo Fisher Scientific) with B27 Supplement minus insulin (A1895601, Thermo Fisher Scientific), 2mM GlutaMAX (35050-061, Thermo Fisher Scientific), and 100mg/mL Penicillin/Streptomycin (30002Cl, Corning). CHIR99021 is a small molecule that activates WNT signaling and initiates the differentiation on day 0 (after the 'day 0' cell collection) (Lian et al. 2012). "Heart media" was replaced 24 hours later at day 1 of differentiation. 48 hours later, at day 3 of differentiation, cells were fed with new "heart media" containing 2uM of the WNT inhibitor Wnt-C59 (5148, Tocris) (Lian et al. 2012). We cultured cells in Wnt-C59 heart media for 48 hours. At day 5, Wnt-C59 was removed and base "heart media" was added. "Heart media" was refreshed on days 7, 10, 12, and 14 of differentiation.

Cells began spontaneous mechanical beating between days 7 and 10 of differentiation (Table S2-1).

Sample Collection and Processing. We performed cardiomyocyte differentiations in batches of two to five cell lines at a time. Every 24 hours from day 0 (iPSC, before treatment with CHIR99021) to day 15 for every cell line, cells in one well of a 6-well culture dish were harvested using mechanical scraping. Cells were rinsed and suspended in PBS and flash-frozen in liquid nitrogen. On day 15 of cardiomyocyte differentiation for all cell lines, we performed flow cytometry to establish purity using a cardiac-specific marker, cardiac Troponin T (564767, BD Biosciences) (Table S2-2). Cells were profiled on the BD LSR-Fortessa Cell Analyzer.

After each time-course was completed, we processed each cell line and balanced our study design with respect to differentiation batch, RNA extraction batch, person who performed the RNA extraction, library batch, and sequencing lane to mitigate technical batch effects (Table S2-1). For all experimental steps after cell collection, all time points of a given cell line were processed together to minimize technical variation related to our factor of interest, which is time. We recorded 27 technical and biological covariates and measured their contribution to variation in our data (Fig. S2-3b).

We extracted RNA from frozen cells using the Qiagen Qiashredder and RNeasy Mini Kit (79656 & 217004, Qiagen). RNA concentration and quality was measured using the Agilent 2100 Bioanalyzer. The average RIN score for all samples was 9.51, with a standard deviation of 1.09.

Library preparation was performed using the Illumina TruSeq RNA Sample Preparation Kit v2 (RS-122-2001 & -2002, Illumina). Libraries in each batch were multiplexed together so

that every sequencing lane contained samples from at least two cell lines. Cell lines were randomized such that lines that were processed together in a sequencing batch were not also together in an RNA extraction batch or a differentiation batch. In total, most sequencing lanes contained 23 to 24 multiplexed samples each. Samples were sequenced 50 base pairs, single-end using the Illumina HiSeq4000 according to manufacturer instructions. The same multiplexed library pool was sequenced twice with the goal of achieving at least 15 million reads per sample (Fig. S2-2).

Genotype data. We used previously collected and imputed genotype data for the 19 Yoruba individuals from the HapMap and 1000 Genomes Project (*86*).

RNA-seq quantification. All RNA-seq samples were aligned to the human genome (GRCh37) using Subread. We counted reads and estimated gene level expression with reads per kilobase million (RPKM) using the `edgeR` R package. We then filtered to genes that were protein-coding, autosomal, and had at least 10 samples such that RPKM >= .1 and raw read counts >= 6. This yielded 16,319 genes. The RPKM distribution in each sample was then quantile normalized and each gene, across all samples, was standardized (mean 0, standard deviation 1).

Biological Replication. We computed replication of day 0 cell lines within previously generated iPSC lines (*32*) and replication of day 15 cell lines within previously generated iPSC-derived cardiomyocyte cell lines (*32*). Notably, the samples from Banovich et al. were also generated in the Gilad lab and use the same panel of iPSCs. Count data from all 4 data sets was re-processed under a uniform pipeline:

1. Count data was $\log_2(\text{count}+1)$ transformed

2. Each gene was standardized to have mean zero and standard deviation 1

3. Top gene expression PCs (in each data set separately) were regressed out.

We regressed out the top 3 PCs in the day 0 and day 15 data sets, top 10 PCs in the Banovich et al iPSC data set, and top 3 PCs in the Banovich et al. iPSC-derived cardiomyocyte data set. The choice of 3 PCs was selected to match the number of PCs in the non-dynamic eQTL analysis. The choice of 10 PCs in the Banovich et al. iPSC data set was selected to match their analysis.

Cell line clustering model (split-GPM). We applied a generative model that assumes a joint clustering over the 19 cell lines and 16,319 genes. That is, the model encodes a global assignment of each of G genes to L gene clusters and assignment of each of N cell lines to K cell line clusters. For each cell line cluster, each gene cluster specifies a Gaussian process (GP) representing a latent gene expression trajectory across time. Thus, the model identifies groups of cell lines with globally different behavior, and groups of genes with similar expression trajectories within each cell line cluster.

Let y_{ng}be the observed gene expression trajectory for gene g in cell line n at times t_{ng}. Our observations are generated as follows:

$$\Phi_n \sim Categorical(\pi)$$

$$\Lambda_g \sim Categorical(\psi)$$

$$f^{kl} \sim GP\big(0, K(\theta)\big)$$

$$y_{ng} |\, \Phi_n = k, \Lambda_g = l, f^{kl}, t_{ng} \sim N(f^{kl}\big(t_{ng}\big), \sigma^2 I)$$

$\pi \in R^K \geq 0\ s.t. \sum_{k=1}^{K} \quad \pi_k = 1,\ \psi \ \in R^L \geq 0\ s.t. \sum_{l=1}^{L} \quad \psi_l = 1$ are cell line cluster mixture weights and gene cluster mixture weights respectively, θ are GP kernel hyperparameters and σ^2

is a global variance parameter. f^{kl} is a function drawn from a gaussian process, while $f^{kl}(t)$ is the function evaluated at points t.

We collect $\{\Phi_n\}_{n=1,\dots N}$ into an NxK binary matrix $\Phi\ s.t.\ \Phi_{nk} = 1 \iff \Phi_n = k$. Likewise, we collect $\{\Lambda_g\}_{g=1\dots G}$ into a GxL binary matrix $s.t.\ \Lambda_{gl} = 1 \iff \lambda_g = l$. The observed data points are conditionally independent given the functions and assignments. Our full likelihood is:

$$p(\{y_{ng}\}|\ \{f^{kl}\}, t_{ng}, \Phi, \Lambda) = \prod_{n,g,k,l}^{N,G,K,L} N(y_{ng}|f^{kl}(t_{ng}), \sigma^2)^{1(\Phi_{nk})1(\Lambda_{gl})}$$

split-GPM approximate inference. Exact computation of the posterior $p(\{f^{kl}\}, \Phi, \Lambda, |\ \{y_{ng}\}, \{t_{ng}\})$ is intractable so we resort to a variational approximation that factorizes and minimizes the KL-divergence of the true posterior:

$$q(\{f^{kl}\}, \Lambda, \Gamma) = \prod_{k,l}^{K,L} q(f^{kl}) \prod_{n}^{N} q(\Phi_n) \prod_{g}^{G} q(\Lambda_g)$$

$$f^{kl} \sim GP(0, K(\theta))$$

$$\Phi_n \sim Categorical(\widehat{\Phi_n})$$

$$\Lambda_g \sim Categorical(\widehat{\Lambda_g})$$

This model bears strong resemblance to the Overlapping Mixture of Gaussian process of Lazaro-Gredilla et.al (*87*) and inference proceeds the same way with the exception that the assignment matrix is decomposed into Φ and Λ. To update the assignments, we iteratively update Φ and Λ until convergence or until a fixed number of iterations is reached.

$$ELBO(q) = E_q[log\ p(\{\{y_{ng}\}|\{f^{kl}\},\{t_{ng}\},\Phi,\Lambda)] + E_q[log\ p(\{f^{kl}\},\Phi,\Lambda)$$
$$- E_q[log\ q(\{f^{kl}\},\Phi,\Lambda)]$$

We iteratively estimate assignment variables and trajectory estimates, then perform gradient based optimization with respect to the kernel parameters. This approximation requires $K \cdot L$ GP regressions, each computed over every data point. To make the problem tractable we further approximate each GP via SVGP (*88*).

In this analysis, we train a model with $K = 2$ cell line clusters, $L = 20$ gene clusters and an RBF kernel with shared length-scale and variance parameters for all $K \cdot L$ clusters.

Non-dynamic cis-eQTL calling per time point. Separately, each time point has a small sample size (maximum of 19 samples). Therefore, we used the WASP combined haplotype test (CHT) (*76*) to increase power, integrating both total expression and allelic imbalance data into the same test, to detect cis-eQTLs in each of the 16 time points, independently. In order to increase accuracy of allele-specific expression estimates, RNA-seq data was re-quantified for eQTL calling by filtering Subread mapped reads using the WASP mapping pipeline under default settings in order to reduce biases in allelic mapping. We tested cis-eQTL association for variants within 50 KB of each gene's transcription start site. Further, we tested the same set of variant-gene pairs in all time points, limiting to variant-gene pairs that passed the following filters in all 16 time points:

1. Variant has minor allele frequency >= .1
2. Gene passes all filters described in "RNA-seq quantification" section
3. Gene has >= 100 reads mapped summed across all cell lines

4. Exon of the gene contains a heterozygous variant in at least 5 cell lines

5. Sum of reads mapping to minor allele across all cell line, heterozygous variant pairs >= 25

These filters yielded 1,009,173 variant-gene pairs (6,362 unique genes) tested in each time point. The same variant-gene pairs were tested in each time point to reduce bias when comparing genetic regulatory effects between time points. We included the first three raw read count expression PCs from samples belonging to the corresponding time point as covariates. The choice to control for three PCs was motivated by maximizing the number of significant non-dynamic eQTLs detected in each time step (Fig. S2-9B). We ran one permutation of the CHT genome-wide. It is worth noting that the CHT is not well calibrated (Fig. S2-10**)**. Multiple testing correction was performed using empirical FDR (eFDR) (*89*) to assess genome-wide significance based on a vector of observed p-values and a vector of null (permuted) p-values. An empirical approach to FDR correction should account and control for the lack of calibration observed when the CHT was applied to our data.

Sparse non-negative matrix factorization. We performed sparse, non-negative matrix factorization of eQTL statistics for all time points to identify broad patterns in eQTL effects. Here, we limited to genes with at least one significant eQTL (eFDR <= .05) across time points. If a gene had more than one significant eQTL, we selected a single variant for that gene with the smallest geometric mean p-value across all 16 time points. We then filled in a matrix, X, where each row represents one gene, each column represents a time point, and each element represents the $-\log_{10}$ p-value corresponding to the row's gene and the column's time point. We then performed sparse non-negative matrix factorization on X (dim NxT) using the python function

`sklearn.decomposition.NMF` (*90*). With K latent factors, this will reduce X into the product of a loadings matrix (L; dim NxK) and a factor matrix (F; dim KxT). F captures shared patterns of eQTL effect sizes across time while L reflects which factors are relevant for each eQTL. All default settings were used except we set `l1_ratio=1` to enforce an element-wise L1 penalty. We ran this analysis for a range of number of latent factors and L1 penalties (alpha) (Fig. S2-11).

Linear dynamic eQTLs. Linear dynamic eQTLs are cis-eQTLs whose effects are linearly modulated by differentiation time. We detected linear dynamic eQTLs with a gaussian linear model that quantified the interaction between genotype and differentiation time on gene expression, while controlling for the linear effects of both genotype and differentiation time. We also controlled for linear effects of the first five cell line collapsed PCs (see below) and, critically, the linear effects of the interaction between the first five cell line collapsed PCs and differentiation time.

We built a separate linear model for each tested variant-gene pair. Specifically, let t denote the time point of the current sample, c denote the cell line of the current sample, T denote the total number of time points, and C denote the total number of samples. $E \in R^{CxT}$ denotes the standardized expression matrix for the current gene, $G \in R^{C}$ denotes the dosage based genotype vector for the current variant, and $PC^{K} \in R^{C}$ denotes the Kth cell line collapsed PC vector. We modeled the expression levels as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \beta_4 PC_c^1 t + \cdots + \beta_{11} PC_c^5 + \beta_{12} PC_c^5 t + \beta_{13} G_c t, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and time (β_{13}). We computed a null distribution by randomly permuting the time point variable that was used for the term capturing the interaction between genotype and time (β_{13}), while keeping the

time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. Using this permutation run, we computed significance with eFDR.

We tested the same set of variant-gene pairs that was tested in the non-dynamic eQTL calling analysis. This was done to reduce bias when comparing non-dynamic eQTLs and dynamic eQTLs.

Cell line confounder estimation using cell line collapsed PCA. Different cell lines can display broadly different patterns of expression across the entire time course, including not only consistent shifts upward or downward in expression of subsets of genes, but different slopes and more generally different expression trajectory shapes (Fig. 2-1B). Variability in slope is of particular concern for detection of dynamic eQTLs – if a subset of cell lines display different slopes over time for many genes, this would lead directly to false positive dynamic eQTLs. Specifically, these cell line subsets reflecting confounders could by chance correspond to the same grouping as genotype across numerous SNPs given the large number of SNPs compared to cell lines. This would then produce apparently large effect $\beta_{13} G_c t$ terms in the dynamic eQTL linear model, and thus numerous false positives. To combat this problem, we used a PCA-based approach we refer to as "cell line collapsed PCA" to identify broad, cell line specific patterns across the entire time course. To do so, we simply rearranged the gene expression matrix from the standard RNA-seq quantification (RPKM levels across 297 samples by 16,319 genes) such that each row was now expression from one cell line and each column was a gene at a single time point. We excluded time points that were not fully observed (days 2, 4, and 13) to avoid missing entries, yielding a final matrix of size 19 by 212,147 (Fig. S2-13). After standardizing each column, we applied PCA to this matrix to learn a low dimensional representation. Here, each cell line has a shared loading across all time points, and PCs reflect trajectories across all

genes, rather than a standard application of PCA with loadings for each sample (a cell line, time point pair).

To ensure that we effectively controlled for the potential confounding effects of cell lines displaying broad trajectory differences over time, we calculated the frequency at which each pair of cell lines share the same genotype across all significant dynamic eQTLs. As noted above, a confounder would cause subsets of cell line to have the same eQTL SNP genotype more often than expected by chance alone, corresponding to cell line clusters with broad differences. In fact, when we do not include cell line collapsed PC loadings in our model, we do see an abundance of such likely false positives (Table S2-4). After controlling for 5 cell line collapsed PCs, the cell lines do not share the same genotype across significant dynamic QTLs more often than background (Fig. S2-16), confirming that cell line PCs help address confounding effects of individual cell line trajectories.

An alternative approach of using pseudo-time, rather than actual time in association testing, does not fully address the problem mentioned here – cell lines don't simply progress faster or slower along the same ultimate trajectory, but seem to deviate in a more complex pattern. Here, this pattern appears to correspond to cell type purity, but more generally, differentiation or any temporal response that follows branching trajectories that can't be captured by a single monotonic pseudo-time term could lead to similar false positives.

We controlled for the first five cell line collapsed PCs and their interaction with differentiation time when detecting both linear and nonlinear dynamic eQTLs. While there does not exist an optimal method to select the number of cell line collapsed PCs, we selected 5 cell line collapsed PCs that: (a) capture most of the variance in gene expression (Fig. S2-14a), (b)

ensure cell lines do not share the same genotype across significant dynamic QTLs more often than background (Fig. S2-16), and (c) result in consistency between non-dynamic eQTLs and dynamic eQTLs (Fig. S2-21 and S2-25).

Simulating expression samples for linear dynamic eQTL power analysis. Using the same notation as defined in the "Linear dynamic eQTLs" section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t + \beta_3 (t * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t, \sigma)$$

For each setting of number of cell lines, t-statistic and minor allele frequency, we simulated 10,000 independent tests (variant-gene pairs) where a specified proportion of those tests follow the null and alternate models. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a specified minor allele frequency. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have the specified minor allele frequency

2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1

3. β_2 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1

4. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1

5. E_{ct} was randomly drawn

6. p-values were computed using the linear model described in the "Linear dynamic eQTLs" section excluding any fixed effects containing cell line collapsed PCs

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Nonlinear dynamic eQTLs. To detect dynamic eQTLs whose effect size changes non-linearly with time, we used a second order polynomial basis function over time, which alters the above linear dynamic eQTL model as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 t^2 + \beta_4 PC_c^1 + \beta_5 PC_c^1 t + \beta_6 PC_c^1 t^2 + \cdots + \beta_{16} PC_c^5 + \beta_{17} PC_c^5 t + \beta_{18} PC_c^5 t^2 + \beta_{19} G_c t + \beta_{20} G_c t^2, \sigma)$$

We quantify the joint effect of the two interaction terms between genotype and time (β_{19} and β_{20}) with a likelihood ratio test with two degrees of freedom using the R `lmtest` package. We computed a null distribution by randomly permuting the time point variable that was used for the two terms capturing the interaction between genotype and time (β_{19} and β_{20}), while keeping the time point variable in all other terms not permuted. An independent permutation was used for every tested variant gene pair. It is worth noting that the nonlinear dynamic eQTLs are not well calibrated (Fig. S2-18**)**. Using this permutation run, we computed significance using eFDR. An empirical approach to FDR correction should account and control for the observed lack of calibration of this test.

Simulating expression samples for nonlinear dynamic eQTL power analysis. Linear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes linearly with

differentiation time. Nonlinear dynamic eQTLs allow us to capture dynamic eQTLs whose effect size changes as a quadratic function of differentiation time. However, both of these approaches are unable to capture arbitrary nonlinear functions of differentiation time. A statistical test that could capture arbitrary nonlinear functions of differentiation time is an ANOVA analysis where time is fit as a factor with 16 levels (ANOVA eQTLs). Here, we simulate several nonlinear dynamic eQTLs and access detection power using three different dynamic eQTL methods:

1. Linear dynamic eQTLs
2. Nonlinear dynamic eQTLs
3. ANOVA dynamic eQTLs

Using a similar notation as defined in the "Linear dynamic eQTLs" section, we define the alternate model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new} + \beta_3 (t_{new} * G_c), \sigma)$$

And the null model as:

$$E_{ct} \sim N(\beta_1 G_c + \beta_2 t_{new}, \sigma)$$

Here, $t_{new}$is a transformation of t. We used four arbitrary transformations of t:

1. $t_{new} = t(t - 10)$
2. $t_{new} = t(t - 7)(t - 15)$
3. $t_{new} = \sin(pi * \frac{t}{5})$
4. $t_{new} = I[t > 7]$

Transformed differentiation time (t_{new}) was scaled to have the same standard deviation as the original values of differentiation time. For each setting of number of cell lines, t-statistic and time transformation, we simulated 10,000 independent tests (variant-gene pairs) where 30% of those tests follow the alternate model and 70% follow the null model. We made the simplifying assumption that each cell line contained 16 time points (T=16). For each test:

1. The genotype vector (G_c) was randomly generated assuming a minor allele frequency of .4. Specifically, both alleles of the variant were drawn independently and both alleles were forced to have a minor allele frequency of .4.

2. β_1 was randomly generated for each test from a separate gaussian distribution with mean 0 and standard deviation of .1

3. β_3 was equal to the t-statistic multiplied by σ. For convenience, σ was fixed to be .1

4. E_{ct} was randomly drawn

5. p-values were computed using the three statistical models described above

Significance of simulated tests was assessed at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data).

Linear dynamic eQTL classifications. We classified the linear dynamic eQTLs as *early* (when the eQTL effect size decreased over time), *late* (when the eQTL effect size increased over time), or *switch* (when the eQTL effect size changes sign over the time course. To do so, we computed predicted eQTL effect size at day 0 and day 15 according to the fitted linear dynamic eQTL model:

Let $\hat{E}_{vg}(t = x, G = y)$ be the predicted expression (according to the fitted dynamic eQTL model) of gene g at time x for a sample with genotype dosage y for variant v. We defined the eQTL effect size ($\beta_{vg}(t = x)$) of variant v on gene g at time x as:

$$\beta_{vg}(t = x) = \hat{E}_{vg}(t = x, G = 0) - \hat{E}_{vg}(t = x, G = 2)$$

If the sign of $\beta_{vg}(t = 0)$ is equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $\left|\beta_{vg}(t = 0)\right| \geq \left|\beta_{vg}(t = 15)\right|$
2. late if $\left|\beta_{vg}(t = 0)\right| < \left|\beta_{vg}(t = 15)\right|$

If the sign of $\beta_{vg}(t = 0)$ is not equal to the sign of $\beta_{vg}(t = 15)$, we assigned that dynamic eQTL to:

1. early if $\left|\beta_{vg}(t = 0)\right| \geq \left|\beta_{vg}(t = 15)\right|$ and $\left|\beta_{vg}(t = 15)\right| <$ thresh
2. late if $\left|\beta_{vg}(t = 0)\right| < \left|\beta_{vg}(t = 15)\right|$ and $\left|\beta_{vg}(t = 0)\right| <$ thresh
3. switch if $\left|\beta_{vg}(t = 0)\right| \geq$ thresh and $\left|\beta_{vg}(t = 15)\right| \geq$ thresh

We assigned thresh = 1.

Nonlinear dynamic eQTL classifications. We classified the nonlinear dynamic eQTLs as early (when the eQTL effect size decreased over time), late (when the eQTL effect size increased over time), switch (when the eQTL effect size changes sign over the time course, or middle (when the eQTL is strongest in the middle of the time course). To do so, we computed predicted eQTL effect size at t=0, t=7.5, and t=15 according to the fitted nonlinear dynamic eQTL model:

$$\beta_{vg}(t=0) = \hat{E}_{vg}(t=0, G=0) - \hat{E}_{vg}(t=0, G=2)$$

$$\beta_{vg}(t=7.5) = \hat{E}_{vg}(t=7.5, G=0) - \hat{E}_{vg}(t=7.5, G=2)$$

$$\beta_{vg}(t=15) = \hat{E}_{vg}(t=15, G=0) - \hat{E}_{vg}(t=15, G=2)$$

If $\beta_{vg}(t=7.5) \geq \beta_{vg}(t=0)$ and $\beta_{vg}(t=7.5) \geq \beta_{vg}(t=15)$, we assigned the dynamic eQTL to middle.

If the sign of $\beta_{vg}(t=0)$ is equal to the sign of $\beta_{vg}(t=15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t=0)| \geq |\beta_{vg}(t=15)|$
2. late if $|\beta_{vg}(t=0)| < |\beta_{vg}(t=15)|$

If the sign of $\beta_{vg}(t=0)$ is not equal to the sign of $\beta_{vg}(t=15)$, we assigned that dynamic eQTL to:

1. early if $|\beta_{vg}(t=0)| \geq |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=15)| <$ thresh
2. late if $|\beta_{vg}(t=0)| < |\beta_{vg}(t=15)|$ and $|\beta_{vg}(t=0)| <$ thresh
3. switch if $|\beta_{vg}(t=0)| \geq$ thresh and $|\beta_{vg}(t=15)| \geq$ thresh

We assigned thresh = 1.

ChromHMM enrichment analysis. We computed enrichment of dynamic eQTLs within cell type specific chromHMM (15 state model) enhancer elements relative to 1,000 sets of randomly selected background variants matched for distance to transcription start site and minor allele

frequency (*77*). We considered the following four chromHMM states to represent enhancer elements:

1. EnhG (state 6)
2. Enh (state 7)
3. BivFlnk (state 11)
4. EnhBiv (state 12)

We used the following five Roadmap cell types to represent iPSCs (*78*):

1. E018: iPS-15b Cells
2. E019: iPS-18 Cells
3. E020: iPS-20b Cells
4. E021: iPS DF 6.9 Cells
5. E022: iPSC DF 19.11 Cells

And the following five Roadmap cell types to represent heart-related cells (*78*):

1. E065: Aorta
2. E083: Fetal heart
3. E095: Left ventricle
4. E104: Right atrium
5. E105: Right Ventricle

To compute enrichment within iPSC specific enhancer elements, we limited to enhancer elements found in at least one of the 5 iPSC cell types and none of the heart-related cell types. Likewise, for enrichment with heart specific enhancer elements, we limited to enhancer elements found in at least one of the 5 heart-related cell types and none of the iPSC related cell types. Odds ratios were smoothed by adding smoothing constant of 1 to each overlap count.

Dilated cardiomyopathy gene set enrichment analysis. We define the dilated cardiomyopathy gene set as the union of all genes in Supplementary Table 3 of Burke et al. 2016. Enrichment was computed via Fisher's exact test.

Supplementary Figures and Tables

Supplementary figures and tables for this chapter are included in Appendix A: Supplementary Figures and Tables.

Chapter III: Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types

Authors:

Katherine Rhodes, Kenneth Barr, Joshua Popp, Benjamin J Strober, Alexis Battle, Yoav Gilad

Abstract

The notion that expression quantitative trait loci (eQTLs) can shed light on disease risk and the associated mechanisms is appealing, and over the past decade there have been many examples to support this scientific premise. Yet, the majority of loci associated with disease risk, though located in putatively regulatory regions, are not currently known to be associated with variation in gene expression. To expose the regulatory role of such loci, we need to characterize gene expression in a large variety of cell types and cellular contexts, most of which are not accessible *in vivo* in humans. To address this challenge, we established a new *in vitro* model system using embryoid bodies (EBs), which are spontaneously differentiating organoids. We generated EBs from induced pluripotent stem cell lines (iPSCs) from three individuals in three replicates, and analyzed single-cell RNA-sequencing data from 42,488 cells harvested from these EBs. Our results show that EBs are composed of dozens of temporally and functionally diverse cell types of all three germ layers. We inferred rooted differentiation trajectories that recapitulate known developmental patterns and discovered gene modules with shared patterns of gene expression dynamics. We show that inter-individual variation contributes significantly to overall patterns of cell type composition, gene expression, and dynamic gene expression. Our results

indicate that EBs can facilitate discovery of QTLs for cell type composition, as well as eQTLs and dynamic eQTLs across a multitude of human cell types and developmental trajectories.

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with human traits and diseases, many of which are located in noncoding regions of the genome and are putatively regulatory in function (66). Gene regulation is dynamic; it varies across cell types and conditions, including environmental exposures and cellular processes such as differentiation (13, 91, 92). To understand regulatory and functional effects of trait-associated variants, it is necessary to perform molecular assays in the relevant cell types at the relevant stages of life (16).

To date, there have been large scale efforts to map genetic variants that regulate gene expression (expression quantitative trait loci, or eQTLs) across human tissues. The GTEx consortium, for example, identified over 4.2 million eQTLs in 54 tissues and cell types (15). Despite this enormous effort, only 11% of disease heritability can be attributed to GTEx cis-eQTLs (14), and only 43% of GWAS variants can be classified as eQTLs (13). Thus, while a meaningful fraction of genetic disease risk can be attributed to effects on gene regulation, the majority of GWAS variants remain unexplained. This gap can be partially attributed to the fact that most studies of human gene regulation are limited to mature, adult tissues that were collected at a single point in time (and almost exclusively, post mortem). It is possible that dynamic and variable regulatory genetic effects, including those that are specific to a given cell type, time point, or environment, may underlie the mechanisms for many unexplained phenotypic associations. For example, recent efforts to characterize gene regulatory dynamics in human induced pluripotent stem cells (iPSCs) and their derived cell types have identified

dynamic eQTLs that are associated with disease risk, supporting the intuitive notion that changes in gene regulation during development may play a significant role in shaping human adult phenotypes, including disease (92, 93).

Human iPSCs offer an ethical and experimentally tractable system for the study of human development and can be differentiated to a wide variety of human cell types. Some directed differentiation protocols result in a single, homogeneous cell type, while others produce heterogeneous populations of cells. In particular, iPSCs can be used to form embryoid bodies (EBs). EBs are three dimensional aggregates of spontaneously and asynchronously differentiating cells; that is, they contain developmentally diverse cell types from all three germ layers. EB formation has been used to verify stem cell pluripotency for decades; yet, until recently, the complexity of EB cellular composition has precluded their use in genomic studies. With single-cell RNA-sequencing (scRNA-seq), it is now possible to characterize the numerous spatially and developmentally distinct cell types within EBs, including transient cell types that would otherwise be inaccessible. Indeed, recent scRNA-seq studies of human EB differentiation have revealed the diversity of cell types composing these structures and the transcriptional dynamics governing early fate decisions (*44*, *94*).

The application of scRNA-seq to EBs should make it possible to map eQTLs in a multitude of cell types arising from the same genotype, including developmental and transient cell types, which are no longer found in adult tissues. The ability to grow multiple cell types in the same dish obviates the need for multiple directed differentiation protocols and affords greater control over confounding variables that might mask genetic effects on gene expression. To date, however, the only studies that have sequenced EB cells have relied on a small sample of cells

from a single individual, leaving a gap in our understanding of technical, biological, and inter-individual variation present in this system (*44*, *94*).

Understanding the sources of variation that affect scRNA-seq data from EBs is crucial for evaluating the utility of EBs as a novel system for eQTL discovery. To this end, we generated EBs from three individuals in three replicates. This allowed us to explore the biological and technical variation present at all levels of this data set; specifically, we evaluated the consistency in cell type composition across replicates and individuals, characterized the structure of variation in gene expression across the entire data set, and finally, captured patterns of dynamic gene expression along distinct developmental trajectories. Our results show that scRNA-seq of differentiating EBs has the potential to be a powerful model system for the study of inter-individual variation in gene regulation across an array of functionally and temporally diverse cell types.

Results

We have performed a pilot study to establish and characterize the EB system. Towards the ultimate goal of performing dynamic eQTL studies using EBs, we have designed a study that allowed us to effectively estimate different sources of variation in single cell data from EBs. In this pilot study, we focused on consistency of the non-directed differentiation process, and the proportion of gene expression variability that can be explained by technical or biological factors.

Study design, data collection, and preprocessing

To characterize sources of variation in gene expression in human EBs, we differentiated EBs from three human iPSC lines in three replicates (see Methods). We performed the experiment in 3 batches, where each batch includes one replicate from each of the 3 individuals.

EBs differentiate quickly, with cell types representing endoderm, mesoderm, and ectoderm present after 8 days (*94*). In this study, we maintained EBs for three weeks after formation, allowing cells to continue to differentiate and mature. After 21 days, we collected scRNA-seq data, targeting equal numbers of cells from each individual in each replicate. After filtering and quality control (Methods), we retained high-quality data from a sample of 42,488 cells (an average of 4,721 cells per individual/replicate). For these cells, we obtained a median of 16,712 UMI counts per cell, which allowed us to measure the expression of a median of 4,274 genes per cell (Fig. S3-1**)**. We integrated data from all cells using *Harmony*, which anchors the data sets by cell type (*95*).

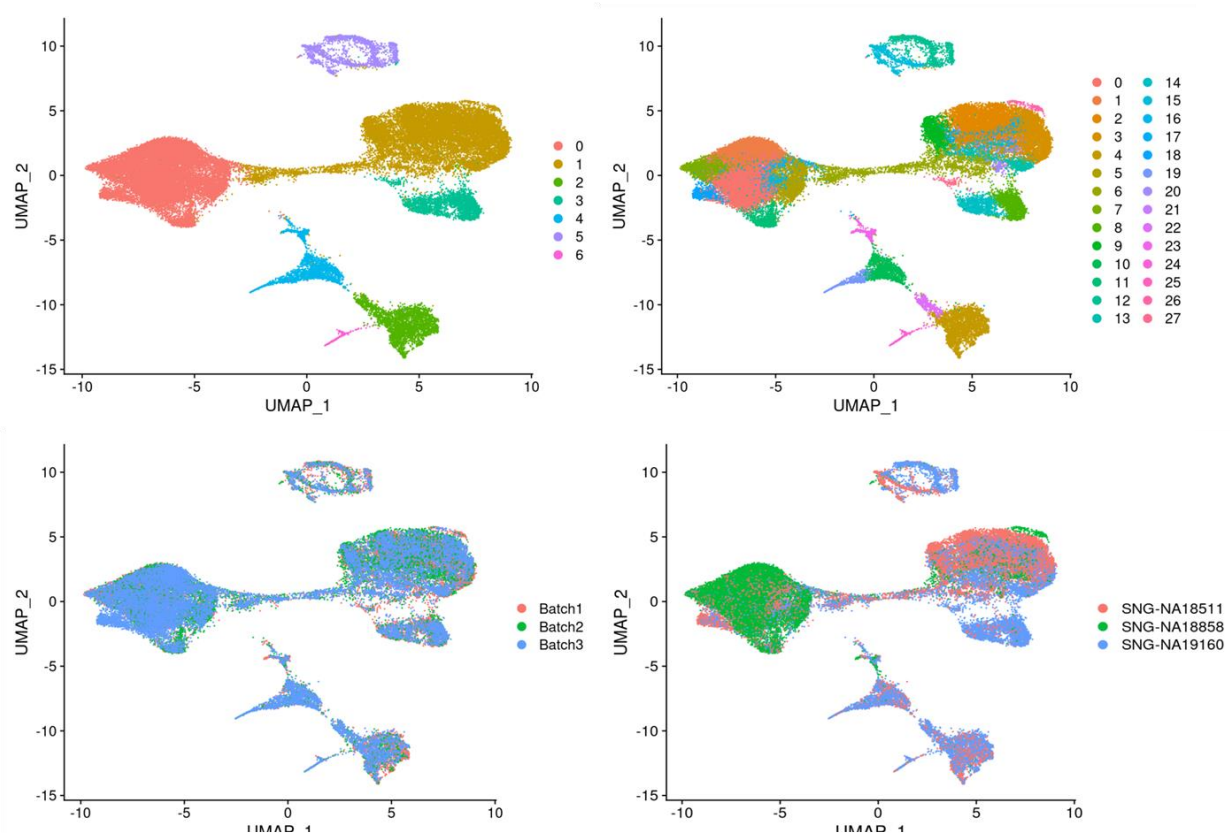


Fig. 3-1. Visualization of cells from EBs with UMAP. TopLeft) Cells are colored by Seurat Cluster assignment at clustering resolution 0.1. TopRight) Cells are colored by Seurat Cluster Assignment at clustering resolution 1. BottomLeft) Cells are colored by replicate BottomRight) cells are colored by individual.

Cell type composition

To validate our expectation that EBs should contain cells from each germ layer, we first characterized the expression of early developmental marker genes. We found cells expressing markers for endoderm (SOX17, FOXA2), mesoderm (HAND1), and ectoderm (PAX6), in addition to cells expressing pluripotency markers (POU5F1, MYC, NANOG)(*96*). We then visualized the data with uniform manifold approximation and projection (UMAP) and observed that cells expressing each of these germ layer markers occupied distinct groups in UMAP space (Fig. S3-2) (*97*). Moreover, we found that every replicate in our experiment, regardless of the individual or batch, includes cells from all three germ layers (Fig. 3-1, Fig. S3-2).

We next sought to further explore the heterogeneous cell types present in these EBs. In studies of scRNA-seq from tissues and samples with well-characterized cell type composition, clustering is often applied to demarcate populations of pure cell types within heterogeneous samples. In these studies, clustering resolution, which determines the number of clusters identified by the algorithm, is typically chosen to recapitulate the expected number of cell types. The identified clusters can be annotated based on the expression of known marker genes. In our case, however, we had no a priori knowledge of the exact number or types of cells that would result from the spontaneous differentiation of the EBs. Hence, we used three complementary approaches to annotate cells, capturing various perspectives on what might define a cell type in this data set. First, we identified cell types by clustering cells and annotating the cell types based on the genes that are highly expressed in each cluster. Second, we annotated cell types by considering the correlation of gene expression in our data with a reference data set of known primary cell types. For our third approach we used a different perspective, and applied topic modeling to consider a less discrete definition of cell type.

For the first approach we used a standard clustering analysis, the Louvain algorithm in Seurat, to identify groups of cells with similar transcriptomes (*98*). To avoid making assumptions about the true number of cell types present, we repeated this analysis across different clustering resolutions (resolution 0.1, 0.5, 0.8, and 1). As expected, the number of clusters we identified varied greatly depending on the resolution; for example, we found seven clusters at resolution 0.1 and 28 clusters at resolution 1 (Fig 3-1) . We performed each subsequent analysis using clusters defined at multiple resolutions, to ensure that our qualitative conclusions are robust to the number of clusters identified.

For each clustering resolution, we calculated pseudobulk gene expression levels using cells from the same cluster, individual, and replicate. To identify marker genes expressed in each cluster, we used *Limma* and *voom* to perform differential expression analysis (Methods) using the pseudobulk estimates. For example, considering the gene expression data of the seven clusters identified at resolution 0.1, we found that the most significantly upregulated genes in each cluster included known marker genes for pluripotent cells (cluster 0), endoderm (cluster 4), mesoderm (cluster 3), early ectoderm (cluster 1), neurons (cluster 5), neural crest (cluster 3), and endothelial cells (cluster 6) (Fig.3-1, Table 3-1). Using this approach at the different cluster resolutions provide a confident set of broad cell type categories present in these data (Table 3-1).

To pursue the second approach, we annotated cells by comparing our gene expression data to available reference sets of scRNA-seq data from primary fetal tissues, human embryonic stems cells (hESCs), and hESC-derived EBs (*99*, *100*). To do this, we first integrated our data set with the reference data sets and visualized cells with UMAP (Fig. 3-2, Fig. S3-3). We observed that reference hESCs cluster closely with pluripotent EB cells. We also observed that the hESC-derived EBs and our iPSC-derived EBs tend to occupy the same areas in UMAP space, implying high overall similarity in cell-type composition despite differing protocols for EB differentiation (and the fact that the experiments were performed in different labs) (Fig. S3-3). EB cells also show overlap with many primary fetal cell types (Fig. 3-2). For example, EB cells annotated as

cluster (res 0.1)	# cells in cluster	top 10 marker genes by logFC	top 10 marker genes by adj. P	Annotation
0	17693	DPPA5, DPPA3, GDF3, NANOG, FGF4, POU5F1, CBR3, PRDM14, DPPA2, TRIML2	TERF1, PHC1, SEPHS1, UGP2, DPPA4, TBC1D23, JARID2, USO1, ZNF398, LRRC47	Pluripotent cells
1	14383	FEZF2, EMX2, LHX2, SOX3, PAX6, WNT7B, ARX, SOX1, ZIC1, SIX3	TPBG, FGFBP3, FZD3, LIX1, SDK2, BTBD17, DACH1, PLAGL1, DEK, ZNF219	Early Ectoderm
2	3086	RGS13, LUM, TECRL, DCN, HAND1, PITX1, COL3A1, SLN, IGF2, FIBIN	TNNI1, COL6A3, COL5A1, RGS4, ACTA2, TMEM88, DOK4, SLC40A1, HAND2, COL3A1	Mesoderm/Stromal Cells
3	2673	MPZ, PRSS56, ROPN1, SOX10, S100B, SCRG1, NPR3, MOXD1, TFAP2B, PHACTR3	NR2F1, CNP, S100B, EDNRA, FGFBP3, ATP1A2, DNAJC1, ZEB2, PHACTR3, METRN	Neural Crest
4	2368	APOA2, CST1, APOA1, APOC3, FGB, RBP4, S100A14, TTR, FGA, APOB	S100A16, LGALS3, GATA3, CST3, KRT19, FN1, EPSTI1, DYNLT3, HDHD3, PKP2	Endoderm
5	1990	NEUROD1, NHLH1, STMN2, NEUROD4, TBR1, STMN4, NEUROG1, SST, ELAVL3, SLC17A6	TAGLN3, RTN1, NHLH1, STMN2, ELAVL2, FNDC5, PCBP4, ELAVL4, DCX, MLLT11	Neurons
6	295	PLVAP, CD34, CD93, CDH5, DIPK2B, PECAM1, EMCN, CRHBP, ESAM, ECSCR	EGFL7, GNG11, RAMP2, IGFBP4, PPM1F, RASGRP3, RCSD1, MAP4K2, PLVAP, DOCK6	Endothelial cells

Table 3-1. Cluster annotation based on differential expression of marker genes.

neural crest based on our gene expression analysis, overlap with primary fetal cell types derived from neural crest like schwann cells and ENS glia (*101, 102*). EB cells annotated as neurons based on our gene expression analysis overlap with fetal neuronal subtypes, including inhibitory neurons, excitatory neurons, granule neurons, ENS neurons, and others. EB cells also show overlap with populations of cells that are rare in the fetal data set like AFP_ALB positive cells (hepatic cells), thymic epithelial cells, and lens fibre cells (Fig. 3-2).

Encouraged by these observations, we expanded the annotation of our EB cells (which until this point were based on the expression of known marker genes) by using the known annotations of the reference primary fetal cell types data set. Specifically, we transferred cell annotations to EB cells based on the nearest neighbor reference cells (Methods) (Fig. S3-4). Using this approach, we found EB cells representing 68 of the 77 cell types present in the reference fetal data set (Fig, S3-4, Table 3-2). The most common annotation was hESC; this can be partially attributed to the high proportion of pluripotent cells in our EB data set, but also to the fact that the reference fetal data set does not include many early development cell types. Indeed, many cells annotated as hESC here are likely to represent immature, differentiating cells which are no longer pluripotent but whose transcriptional profiles more closely match hESCs than the more highly differentiated fetal cell types present in the reference data set. In this sense, EB data sets may capture transient developmental cell types that are difficult or impossible to study even in fetal primary samples. Outside the hESCs, many fetal cell types are only represented by small populations of EB cells. For example, only 1 EB cell is annotated as a sympathoblast, and only 2

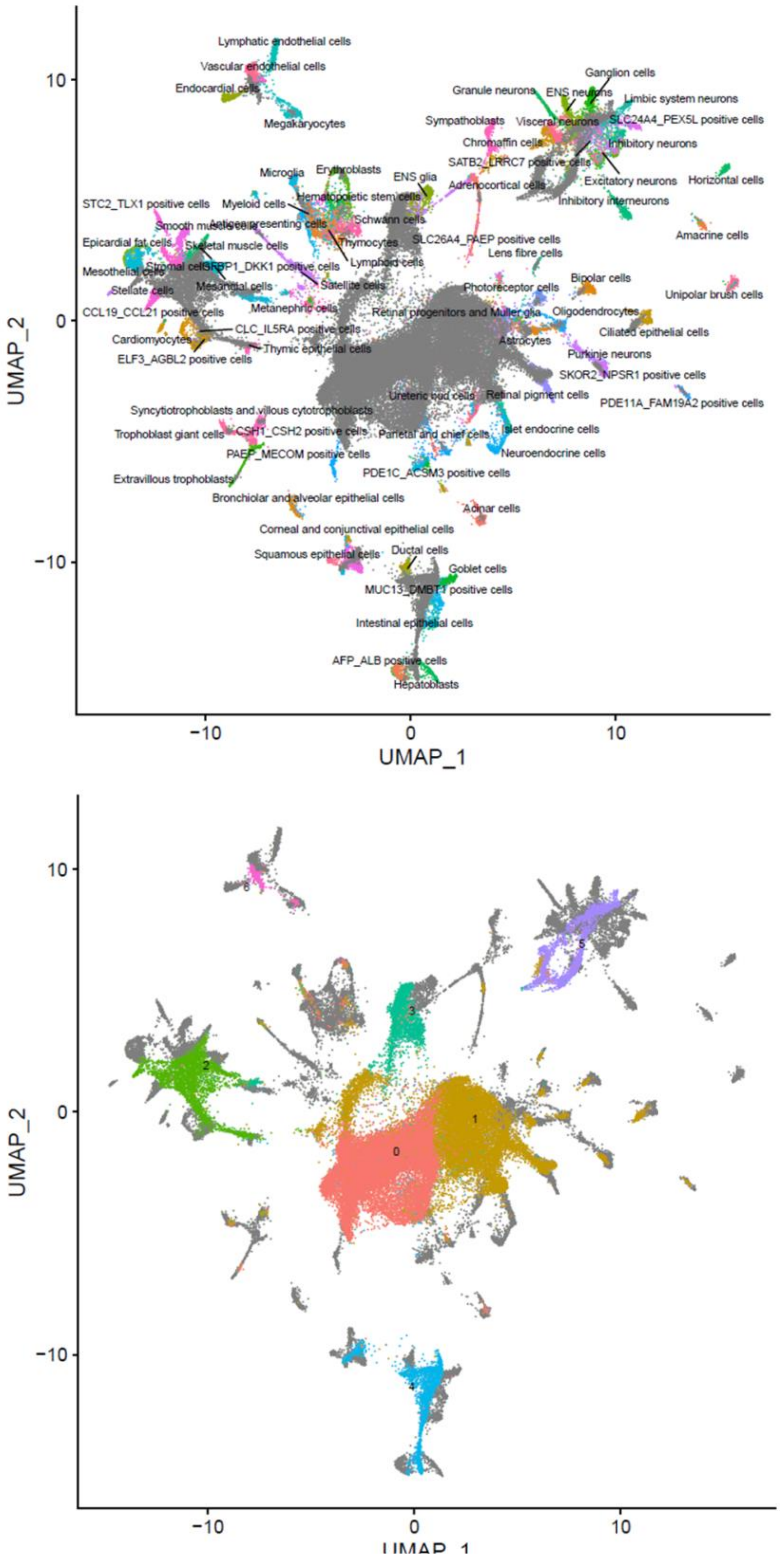


Fig. 3-2. UMAP visualization of EBs and integrated fetal reference data. Left) Cells are colored by cell types present in the Cao et al. data set, with grey points representing EB cells. Right) Cells are colored by Seurat cluster identity at clustering resolution 0.1, with grey points representing cells from the Cao et al. reference set.

cells annotated as thymocytes. This observation indicates that we need a deeper sampling of our EBs in order to properly explore their true cell type composition. Overall, annotation based on the reference set revealed the presence of dozens of diverse cell types in EBs.

Topic modeling of the single cell gene expression data

Both of the approaches we described above, clustering, and comparison to a reference data set, assume that "cell types" are discrete categories. Accordingly, each cell has a single true identity and the cell type categories are assumed to be homogeneous and static. This definition of

Cell Type Annotation	Frequency in EB cells
Acinar cells	52
Adrenocortical cells	7
AFP_ALB positive cells	36
Amacrine cells	79
Antigen presenting cells	1
Astrocytes	39
Bipolar cells	40
Cardiomyocytes	521
CCL19_CCL21 positive cells	2
Chromaffin cells	24
Ciliated epithelial cells	314
CLC_IL5RA positive cells	3
Corneal and conjunctival epithelial cells	21
Ductal cells	329
ELF3_AGBL2 positive cells	23
Endocardial cells	5
ENS glia	30
ENS neurons	72
Epicardial fat cells	112
Erythroblasts	34
Excitatory neurons	1
Ganglion cells	61
Goblet cells	271
Granule neurons	158
Hematopoietic stem cells	9
Hepatoblasts	212
Horizontal cells	53
IGFBP1_DKK1 positive cells	341
Inhibitory interneurons	9
Inhibitory neurons	65
Intestinal epithelial cells	501
Islet endocrine cells	552
Lens fibre cells	51
Limbic system neurons	15

Cell Type Annotation	Frequency in EB cells
Lymphatic endothelial cells	2
Megakaryocytes	37
Mesangial cells	222
Mesothelial cells	160
Metanephric cells	142
Microglia	44
MUC13_DMBT1 positive cells	181
Myeloid cells	1
Neuroendocrine cells	4
Oligodendrocytes	53
Parietal and chief cells	14
PDE11A_FAM19A2 positive cells	35
Photoreceptor cells	9
Purkinje neurons	35
Retinal pigment cells	391
Retinal progenitors and Muller glia	21
scHCL.hESC	34086
Schwann cells	38
Skeletal muscle cells	5
SKOR2_NPSR1 positive cells	39
SLC24A4_PEX5L positive cells	297
Smooth muscle cells	60
Squamous epithelial cells	160
Stellate cells	247
Stromal cells	217
Syncytiotrophoblasts and villous cytotrophoblasts	17
Thymic epithelial cells	116
Thymocytes	1
Trophoblast giant cells	5
uncertain	1566
Unipolar brush cells	37
Ureteric bud cells	26
Vascular endothelial cells	58
Visceral neurons	119

Table 3-2. Frequency of cell type annotations among EB cells. Frequencies are based on annotations transferred from the 5 most similar references cells. If cells could not be confidently annotated as a particular reference cell type, they are annotated as 'uncertain'.

a cell type is intuitive and is often practical for genomic studies, making it possible to consider results from single cell analysis in the context of the wealth of knowledge previously gained from bulk assays. However, partitioning cells into discrete groups is unlikely to capture the full degree of complexity in gene expression heterogeneity of single cells, particularly in a data set that includes differentiating cells, which are expected to have varying degrees of similarity to a terminal cell type. In an alternate paradigm, cell type can be viewed as continuous, with the expression profile of each cell representing grades of membership to discrete categories (*103*). One method used to capture cell identity in this paradigm is topic modeling, which learns major patterns in gene expression within the data set, or topics, and models each cell as a combination of these topics.

We applied topic modeling using *fastTopics* at a range of topic resolutions, identifying 6, 10, 15, 25, and 30 topics in our data (*103*, *104*). Some topics correspond closely to Seurat

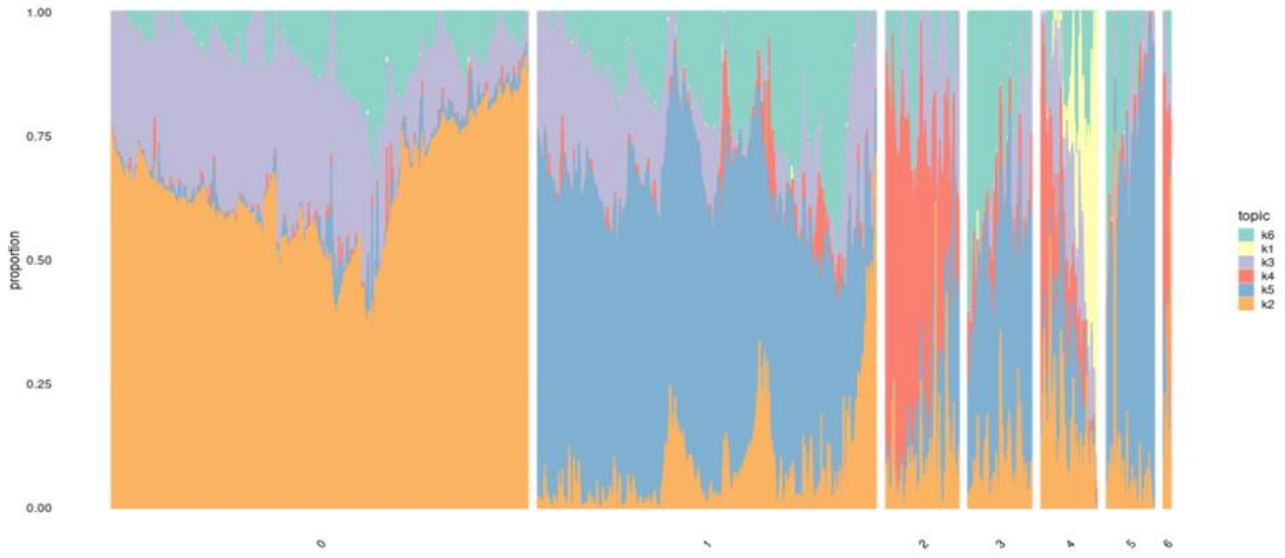


Fig. 3-3. Structure plot showing the results of topic modelling at k=6. Structure plot of a random subset of 5000 cells. Each column represents a single cell colored by the loading of each topic. Plot is divided by Seurat cluster at resolution 0.1.

clusters, showing loading on cells of that cluster but not on others. For example, in the k=6 topic analysis, topic 1 is loaded exclusively on cells assigned to Seurat cluster 4 (cluster resolution 0.1) which we previously annotated as endoderm (Fig.3-3, Table 3-1). Compared to other topics, topic 1 shows significant increase in expression of FN1, AFP, and GPC3, which are known markers of hepatocytes (Fig. S3-5, Table S3-2). Seurat clustering at higher resolution (resolution 1) results in further categorical division of this large endoderm group of cells into definitive endoderm and hepatocytes. Using topic modeling, however, we found that cells, in fact, have a gradient of membership in topic 1, likely capturing different temporal stages of hepatocyte differentiation.

Certain topics are shared across cells assigned to different Seurat clusters (Fig. S3-6). For example, topic 6 from the k=6 topic analysis is loaded across all Seurat clusters; compared to all other topics, topic 6 shows increased expression of many ribosomal genes, housekeeping genes (GAPDH), and genes coding for proteins involved in cellular metabolism (LDHA) (Table S3-2). This indicates that topic 6 captures patterns of gene expression associated with cellular processes and functions that are not specific to a particular cell type. This again highlights an advantage of topic modeling, enabling us to explore variation in the representation of gene expression profiles associated with processes shared across broad cell types.

Biological and technical sources of variation

Once we functionally annotated the cells using the three approaches we discussed, we sought to understand the consistency in cell type composition across individuals and between replicates. We began by calculating the proportion of cells that were assigned to each Seurat cluster at resolution 0.1 for each replicate. We then performed hierarchical clustering of the samples based on the proportion of cells in each Seurat cluster (Fig. S3-7). Using this approach,

replicates of 18858 cluster together. Among individuals 18511 and 19160, there is a more complex sub-clustering. Replicates 1 and 2 of 18511 cluster together while replicate 1 of 19160 clusters away from the others. There is, however, no distinct clustering by replicate, suggesting that the observed pattern is due to the overall high degree of similarity in cell type composition of all replicates of these two lines.

We repeated this analysis at a range of cluster resolutions and determined that this finding is robust with respect to the number of clusters; in particular, at the highest resolution, using Seurat clustering resolution of 0.5 and 1, the samples cluster perfectly by individual (Fig. S3-7). We also repeated this analysis using topic loadings as a measure of cell type composition. We calculated the loading of each topic on each individual-replicate group and performed hierarchical clustering (Fig. S3-8). Again, we found that at varying values of k, samples generally cluster by individuals, but using the higher resolution topic-based approach we also observed substantial variation between replicates (Fig. S3-8).

One individual in particular (18858) always clusters away from the other two lines, showing a distinct distribution of cell types that is consistent across all replicates. The EBs from this individual have a particularly high proportion of pluripotent cells and a lower differentiation efficiency compared to the other two lines. A previous study of iPSC-derived dopaminergic neurons from Jerber et al. found that patterns of gene expression in iPSC subpopulations could predict differentiation efficiency (*105*). This study found that the existence of a subpopulation of iPSCs with increased expression of UTF1 was associated with poor differentiation efficiency (Fig. S3-9). This is likely to be due to cell line intrinsic factors rather than genetic variation. Indeed, we found that UTF1 is significantly differentially expressed between individuals in the pluripotent cluster, with higher expression in 18858. This suggests that the relationship between

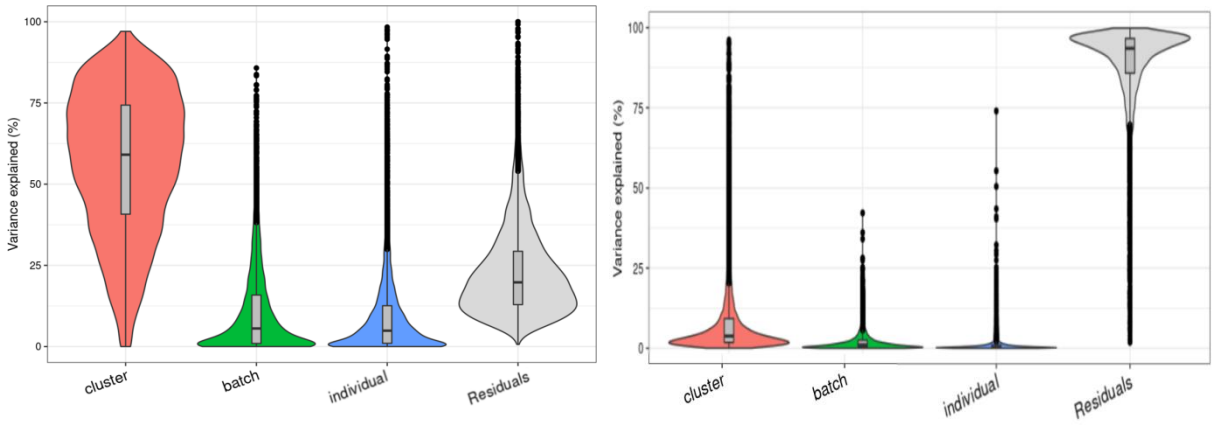


Fig. 3-4. Percent of gene expression variance explained by biological and technical factors. Violin plot showing the percent of variance in gene expression explained by cluster, replicate (batch), and individual in this data set after partitioning variance in pseudobulk samples (left) and at single cell resolution (right).

UTF1 expression and neuronal differentiation efficiency identified by Jerber et al. may apply broadly to differentiation efficiency.

We further characterized determinants of variation in our system by considering factors that contribute to variation in gene expression levels. Hierarchical clustering of pseudo-bulk expression estimates from each sample shows that, as might be expected, samples tend to cluster first by cell type (Seurat cluster), then by individual and replicate (Fig. S3-10). We performed variance partitioning using pseudo-bulk expression levels of cells from the same cluster, replicate, and individual to estimate the relative contribution of cell type, individual, and replicate to overall patterns of gene expression variation (Fig. 3-4) (*106*). We found that cell type identity explained the largest proportion of variation at all clustering resolutions tested (variance explained median value ~60% at clustering resolution 0.1) and that replicate and individual explained approximately equal proportions of the variance (each explains a median value of ~5% of variance). Depending on clustering resolution, a median value of approximately 20-30% of variance is explained by residuals, which can be attributed to noise or other technical variation

not captured in the model. We then partitioned the variance in gene expression at single cell resolution (instead of using pseudo-bulk estimates) and found that cluster explains more variation on average than replicate and individual, and that replicate explains more variation on average than individual (Fig. 3-4). At single cell resolution, residuals explain a median value of 93% of the variation, which is expected due to the high degree of variance in gene expression profiles among individual cells.

To determine whether biological and technical factors contributed differently to variation between cell types, we also partitioned the variance due to replicate and individual in each Seurat cluster separately (Fig. S3-11). The results are not uniform across clusters. At clustering resolution 0.1, individual contributes more to variation, on average, in clusters 0, 1, 4, and 5, while batch contributes more to variation in clusters 2, 3, and 6. Notably, clusters 2, 3, and 6 include only few cells from individual 18858 (Table S3-3). Studies that incorporate a larger number of cells will increase representation of rare cell types, which will increase power to study patterns of gene regulation. In every cluster, variation due to replicate dominates the variation of certain genes but not others. This complex structure indicates that, unlike most other eQTL studies, future studies of EBs need to implement study designs with multiple replicates to appropriately account for this variation.

Dynamic patterns of gene expression

Arguably, the most attractive property of single cell data from the EB system is the ability to study dynamic gene regulatory patterns throughout differentiation. In order to explore dynamic patterns of gene expression, we inferred developmental trajectories using PAGA (*107*). The coarse-grained PAGA graph shows edges that represent likely connections between cell clusters (clustering resolution 1) and we were able to trace developmental trajectories through

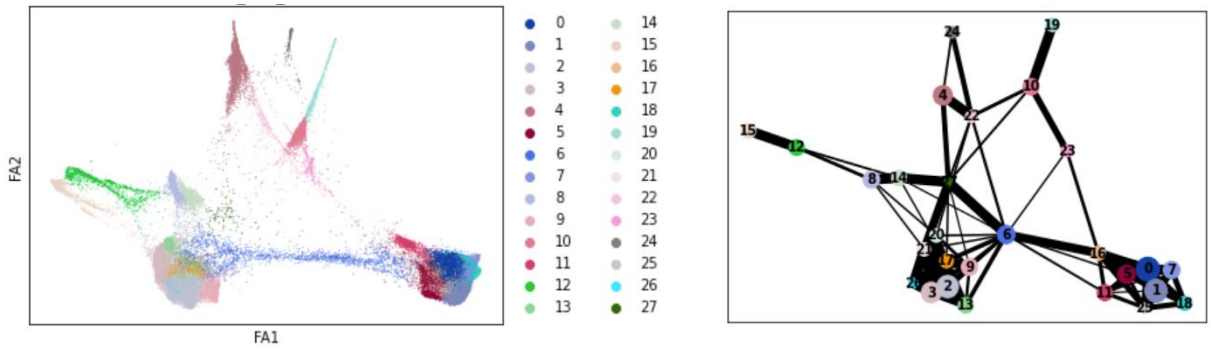


Fig. 3-5. Force Atlas and Coarse-Grained PAGA graph. Left) Force Atlas plot showing cells colored by seurat cluster at resolution 1. Right) Coarse-Grained PAGA graph where each node represents a seurat cluster (clustering resolution 1) and edges represent inferred connections between clusters, with the edge weight corresponding to the statistical strength of the connection.

this path (Fig. 3-5). Since the EBs still include undifferentiated pluripotent cells, we were able to define rooted trajectories beginning at the known starting point. Using this approach we inferred differentiation branches to each germ layer originating from cluster 22, which expresses primitive streak markers (MIXL1, EOMES), showing recapitulation of developmental trajectories defined during gastrulation (*108*). We also note that hepatocytes (cluster 19), an endoderm-derived cell type, branch off of the endoderm cluster (cluster 10). Endothelial cells (cluster 24), which are derived from mesoderm, branch off from the mesoderm cluster (cluster 4), and neurons (clusters 12, 15), an ectoderm-derived cell type, branch off from the early ectoderm clusters (clusters 2, 3, 8, 9, 13, 14, 17, 20, 21, 26, and 27) (Fig. 3-5). We then assigned pseudotime values to each cell using the diffusion pseudotime method with pluripotent cells (cluster 1) defined as the root (*109*).

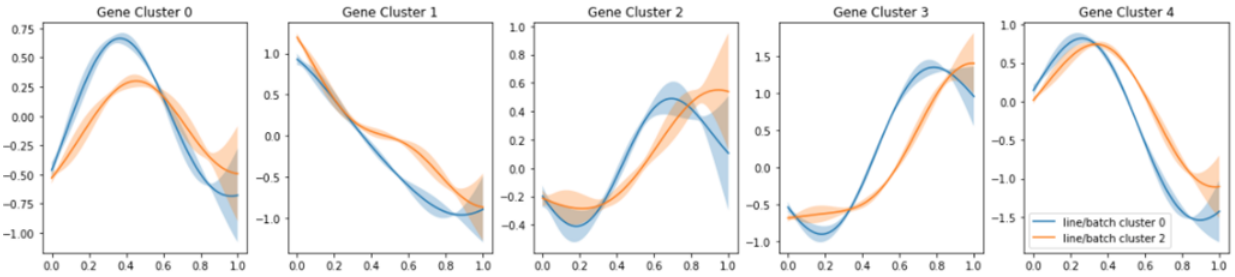
We manually traced high confidence trajectories through the data representing the progression from pluripotent cells to hepatocytes (clusters 0, 1, 5, 6, 7, 10, 11, 16, 18, 19, 25, and 22), pluripotent cells to endothelial cells (clusters 0, 1, 4, 5, 6, 7, 11, 16, 18, 22, 24, and 25), and

pluripotent cells to neurons (clusters 0, 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 15, 26, and 27) (Fig. 3-5). For groups of clusters with a higher degree of connectivity (e.g. clusters expressing pluripotent markers and clusters expressing early ectoderm markers), all clusters within the region with high connectivity were included in the trajectory to avoid choosing an arbitrary path through these clusters. Next, we applied split-GPM, an unsupervised probabilistic model, to infer dynamic patterns of gene expression within a particular developmental trajectory, while simultaneously performing clustering of genes and samples. Split-GPM is built for use with time course, bulk RNA-seq data; therefore, we calculated pseudo-bulk expression values for individual-replicate groups within decile bins of pseudo-time. We were able to identify gene modules with distinct dynamic patterns of expression along the trajectories to neurons, hepatocytes, and endothelial cells (Fig.3-6, Fig. S3-12, Fig. S3-13).

Gene set enrichment analysis of these modules shows expected dynamic patterns for certain gene sets. For example, we found that gene modules that increase expression through pseudo-time along the differentiation trajectory to hepatocytes, which are the predominant cell type of the liver and are responsible for the production of bile, are enriched for the hallmark bile acid metabolism and fatty acid metabolism gene sets (Fig. 3-6). In the trajectory leading to endothelial cells, which are derived from mesoderm, we found that a gene module with high expression at intermediate pseudo-time values is enriched for hallmark genes expressed during the epithelial-mesenchymal transition, which is essential for mesoderm formation (Fig. S3-12) (*110*). In all three trajectories, gene modules characterized with higher expression at low pseudo-time values show enrichment for gene sets related to the cell cycle (Fig.3-6, Fig.S3-12, Fig.S3-13); this is expected because pluripotent cells at the lowest pseudo-time values tend to grow and

divide faster than more differentiated and more mature cell types, which often exit the cell cycle (*111*).

To determine the consistency in dynamic patterns of gene expression between replicates and individuals, we ran split-GPM ten times on cells from the neuron, hepatocyte, and endothelial cell lineages and observed how often each pair of individual-replicate samples clustered together (Fig. 3-7)(*92*). All three replicates of 18511 and two replicates of 19160



	1	3	4
	bonferonni-adjusted	bonferonni-adjusted	bonferonni-adjusted
HALLMARK_CHOLESTEROL_HOMEOSTASIS	1	2.11447e-06	1
HALLMARK_MITOTIC_SPINDLE	1	1	1.10513e-09
HALLMARK_G2M_CHECKPOINT	1	1	1.33349e-20
HALLMARK_ADIPOGENESIS	1	2.20281e-05	1
HALLMARK_ANDROGEN_RESPONSE	1	0.00384152	1
HALLMARK_COMPLEMENT	1	0.000290349	1
HALLMARK_MTORC1_SIGNALING	0.00367589	1	1
HALLMARK_E2F_TARGETS	1	1	5.70599e-20
HALLMARK_MYC_TARGETS_V1	1	1	7.52984e-08
HALLMARK_MYC_TARGETS_V2	0.000390504	1	1
HALLMARK_XENOBIOTIC_METABOLISM	1	3.47546e-07	1
HALLMARK_FATTY_ACID_METABOLISM	1	6.14727e-06	1
HALLMARK_COAGULATION	1	1.62817e-15	1
HALLMARK_BILE_ACID_METABOLISM	1	1.1601e-07	1
HALLMARK_PEROXISOME	1	0.000343498	1

	cell line/batch	assignment
0	SNG-NA18511--Batch1	0
1	SNG-NA18511--Batch2	0
2	SNG-NA18511--Batch3	2
3	SNG-NA18858--Batch1	2
4	SNG-NA18858--Batch2	2
5	SNG-NA18858--Batch3	2
6	SNG-NA19160--Batch1	0
7	SNG-NA19160--Batch2	0
8	SNG-NA19160--Batch3	2

Fig. 3-6. Split-GPM clustering within the hepatocyte trajectory. Top) Table showing bonferonni-adjusted p-values from gene set enrichment analysis of gene modules. Bottom Left) Dynamic expression patterns of identified gene modules in each cluster of replicate-individual samples. Bottom Right) cluster assignments of each individual-batch sample based on shared patterns of dynamic gene expression

always cluster together, indicating that these two lines share the same expression dynamics overall, although the first replicate of 19160 seems to have a distinct pattern of dynamic expression and often clusters separately. We again observed that replicates of 18858 show patterns distinct from the other two lines as well as greater variation between replicates, likely

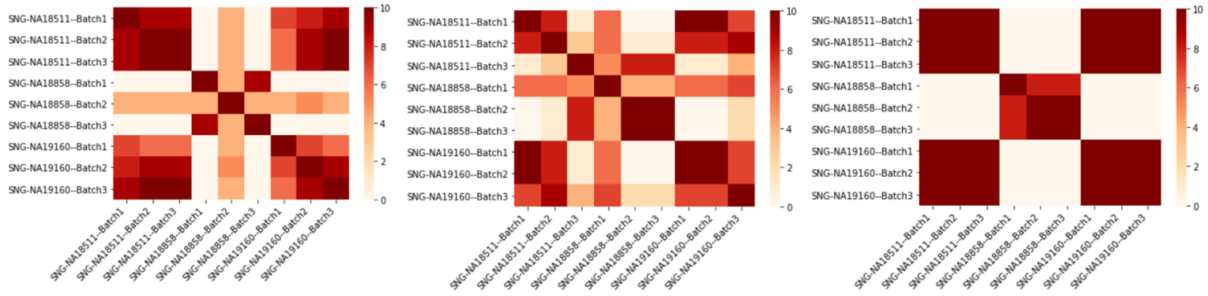


Fig. 3-7. Heatmaps showing frequency of Split-GPM cluster assignments. Heatmaps show the frequency with which batch-individual groups were assigned to the same cluster after running split-GPM in a certain trajectory 10 times. Left) results from the neural trajectory. Center) results from the hepatocyte trajectory. Right) results from the endothelial trajectory

due to the overall relative poor differentiation efficiency of this line. In the endothelial lineage, clustering shows similar patterns seen in the neuronal lineage (Fig. 3-7). All replicates of 18511 and 19160 cluster together, indicating that these lines share dynamic expression patterns in the endothelial lineage as well. 18858 replicates almost always clustered together, but never clustered with the other two individuals, indicating that this line's dynamic expression patterns in the endothelial lineage are consistent but distinct. In the hepatocyte lineage, we observed stronger replicate-specific differences (Fig. 3-7). Replicates of individual 19160 still tend to cluster together and to cluster with replicates 1 and 2 of 18511. Replicate 3 of 18511 never clusters with the other replicates of that individual. Replicate 1 of 18858 also shows patterns distinct from the other two replicates of that individual, indicating that there were replicate-specific effects on dynamic gene expression.

Discussion

This work represents a thorough exploration of heterogeneity in single cell data obtained from human EBs towards the goal of establishing this system as a tool to enable studies of variation in human gene regulation across a range of spatially and temporally diverse cell types. We used iPSC-derived embryoid bodies because this *in vitro* model system circumvents the logistical challenges and ethical barriers associated with studies of primary human developmental tissues. This system has key advantages over studies of primary tissues; for example, we are able to control cellular environment as well as biological factors including age, sex, and ancestry. Further, we can generate EBs comprised of the same set of diverse cell types from large samples of individuals, enabling high-powered comparisons of cell-type specific gene expression.

In subsequent studies we plan to leverage embryoid bodies to identify QTLs and dynamic QTLs across diverse terminal and differentiating cell types. This, of course, raises an ostensibly critical question: to what extent do the cell types derived from embryoid bodies faithfully model immature, developing cells *in vivo*? There is no doubt that the *in vitro* EB differentiated cells are not a perfect model of primary cell types. The question is whether EB cells are sufficiently representative of primary cell types to be useful. To address this question, we performed several analysis, which suggest that the EB mode can be reasonable useful. Specifically, we found that EB cell types express known cell-type specific marker genes, including markers of known developmental stages. EB cells also cluster with more than 70 diverse primary cell types from a reference panel of fetal tissues and hESCs, including rare fetal cell types. Lastly, we identified gene modules with dynamic expression patterns that match broad expectations of developmental biology. Together, these results provide evidence that EB cell types are a suitable model of both terminal and developmental cell types.

Moreover, EBs may be a useful model for understanding the genetic underpinnings of human traits and disease regardless of the degree to which they faithfully model human development. EB-derived cells represent a wealth of previously unstudied cell states and dynamic processes. Hypothetically, QTLs identified in these cell types still represent biologically meaningful differences in genetic control of gene regulation, whether they manifest in human development or in adult tissues upon a particular environmental exposure. To provide an anecdotal example of this reasoning, we considered previously collected data from *in vitro* differentiation experiment. We took a closer look at the 28 middle-dynamic eQTLs Strober et al. identified during the differentiation of iPSCs to cardiomyocytes (*92*). Middle-dynamic eQTLs have their strongest genetic effect at intermediate stages of the differentiation time course, and most of them (25/28) were only identified in intermediate stages of differentiation. This means that these eQTLs are active in early *in vitro* differentiating cells whose fidelity to primary developing cell types has not been ascertained. These 28 dynamic eQTLs were entirely novel and had not been identified as cis eQTLs in any tissue in the GTEx data set. Strober et al. reported that one of these middle-dynamic eQTLs was also found to overlap a GWAS variant associated with body mass index and red blood cell count. This finding highlights that dynamic eQTLs acting in early cell types in *in vitro* differentiations may affect long-term disease risk in adults.

To further explore the utility of dynamic eQTLs identified in *in vitro* differentiations, we used GTEx data to ask whether the middle-dynamic eQTLs are associated with inter-individual variation in trans gene expression and/or cell composition. Trans eQTL associations are more tissue-specific than cis eQTLs, but trans eQTLs are much harder to identify because of their small effect sizes and the requirement to test the association of every locus with every gene. We

identified a middle dynamic eQTL SNP (rs6700162) that is associated with fibroblast cell type proportions in HLV (heart left ventricle; $p < 0.0009$) and with cardiac muscle cell proportions in HLV ($p < 0.003$). This SNP was also found to have a trans eQTL p-value of 1x10-5 in coronary artery. Without the prior knowledge provided by dynamic eQTL data from the *in vitro* differentiated cardiomyocytes, it would have been impossible to identify these associations using adult primary tissues because the burden of multiple testing when the entire GTEx data set is considered is prohibitively large. This example implies that developing EB cells could be used to understand how transient effects on gene expression are propagated into functional, long-lasting consequences downstream.

Summary

Human embryoid bodies have the potential to be a powerful system for the identification of dynamic eQTLs. In this pilot study, we performed foundational analyses to better understand how to appropriately conceptualize heterogeneity in this kind of data and how to best design large-scale studies of embryoid bodies. In this pilot study, we explored cell type composition of embryoid bodies in two paradigms; first, with discrete cell types identified with a traditional clustering algorithm, then with more continuous cell "types" identified with topic modeling. Cell types defined by discrete clustering are often easier to interpret because they can be contextualized with marker genes and reference data sets defined with bulk sequencing. We conclude, however, that topic modeling is more appropriate for highly heterogeneous single cell data sets like this one. We also explored sources of variation in cell type composition and gene expression. We found that individual variation primarily contributes to patterns in cell type composition based on both discrete clustering and topic modeling. However, variation between replicates is non-negligible, indicating that future studies should focus on inter-individual

variation in cell type composition. We also found that technical variation between replicates contributes significantly to variation in gene expression. Future efforts to map regulatory QTLs in EBs should implement study designs with multiple replicates to appropriately correct for batch effects. Overall, this pilot study has laid the groundwork to transform embryoid bodies into a powerful model system for the understanding of human gene regulation.

Methods

Samples

We used iPSC lines from three unrelated individuals from the Yoruba HapMap population to form EBs. The iPSC lines were reprogrammed from lymphoblastoid cell lines and were characterized previously (*32*). Two of the lines (18511, 18858) are from female individuals and one (19160) is from a male individual.

iPSC maintenance

We maintained feeder-free iPSC cultures on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with StemFlex Medium (A3349401, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). We grew cells in an incubator at 37°C, 5% CO2, and atmospheric O2. Every 3-5 days thereafter, we passaged cells to a new dish using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded cells with ROCK inhibitor Y-27632 (ab120129, Abcam).

Embryoid body formation and maintenance

We formed EBs using a modified version of the STEMCELL Aggrewell400 protocol. Briefly, we coated wells of an Aggrewell 400 24-well plate (34415, STEMCELL) with anti-

adherence rinsing solution (07010, STEMCELL). We dissociated iPSCs and seeded them into the Aggrewell400 24-well plate at a density of 1,000 cells per microwell (1.2x106 cells per well) in Aggrewell EB Formation Medium (05893, STEMCELL). After 24 hours, we replaced half of the spent media with fresh Aggrewell EB Formation Medium. 48 hours after seeding the Aggrewell plate, we harvested EBs and moved them to an ultra-low attachment 6-well plate (CLS3471-24EA, Sigma) in E6 media (A1516401, ThermoFisher Scientific). We maintained EBs in culture for an additional 19 days, replacing media with fresh E6 every 48 hours.

Embryoid body dissociation

We collected and dissociated EBs 21 days after formation. We dissociated EBs by washing them with phosphate-buffered saline (PBS) (Corning 21-040-CV), treating them with AccuMax (STEMCELL 7921) and incubating them at at 37°C in for 15-35 minutes. After 10 minutes in Accumax, we pipetted EBs up and down with a clipped p1000 pipette tip for 30 seconds. We repeated pipetting every five minutes until EBs were completely dissociated. We then stopped dissociation by adding E6 media and straining cells through a 40um strainer (Fisherbrand 22-363-547). We resuspended cells in PBS and counted them with a TC20 Automated Cell Counter (450102, BioRad). Before scRNA-seq, we mixed together an equal number of cells from each line.

Single cell sequencing

We collected scRNA-seq data using the 10X Genomics V3.0 kit. Single cell collections for this experiment were mixed with cells from a larger experiment in all three replicates. From the first replicate of EB differentiations, we mixed YRI cells with EB cells from an additional three humans and chimpanzees (9 individuals total). Even numbers of cells from all 9 individuals, cells were collected across 9 lanes of a 10x chip, targeting 10,000 cells per lane. In

this replicate, reagents from 3 different 10x kits were used. From replicates 2 and 3 of EB differentiation, EBs were only generated from the YRI individuals and the 3 chimpanzees (6 individuals total). In each replicate, we mixed even numbers of cells of each individual and collected cells in 4 lanes of a 10x chip, targeting 10,000 cells per lane, and samples were processed using reagents from a single 10x kit.

Libraries were sequenced using paired-end 100 base pair sequencing on the HiSeq 4000 in the University of Chicago Functional Genomics Core. For libraries from replicate 1, we mixed equal proportions of each of the 6 out of the 9 libraries and sequenced the pooled samples on 1 lane of the HiSeq 4000. Preliminary analyses showed that 2 of these lines were low quality. We remade one of the low quality libraries and discarded the other. We then mixed equal proportions of the remade library with the remaining 3 libraries from replicate 1 and sequenced the pooled samples on one lane of the HiSeq 4000. Preliminary analyses indicated that 3 out of 4 of these libraries were below optimal quality, but would produce usable data. We then pooled together samples from the final 8 libraries from replicate 1, mixing equal parts of each of the 5 high quality libraries with half the amount of the other 3, and deep-sequenced this pool on 8 lanes of the HiSeq 4000. For replicate 2 libraries, we mixed equal parts of all 4 libraries and sequenced on 1 lane. After confirming that each library was high quality, we deep-sequenced the same pool on 6 additional lanes of the HiSeq 4000. For replicate 3 libraries, we mixed equal parts of all 4 libraries and sequenced on 1 lane. After confirming that each library was high quality, we deep-sequenced the same pool on 4 additional lanes of the HiSeq 4000. In all cases, the number of lanes for deep sequencing was calculated to reach 50 % saturation.

Alignment and sample deconvolution

We used STARsolo to align samples to human protein coding genes (GRCh38) (*112*). We demultiplexed individual samples and identified doublets using demuxlet (*113*). For this demultiplexing with demuxlet, we used previously collected and imputed genotype data for these three Yoruba individuals from the HapMap and 1000 Genomes Project and as well as genotype data from the other human and chimpanzee individuals included in the sequenced libraries (*114*, *115*).

Filtering and integration

We ran EmptyDrops to identify barcodes tagging empty droplets and kept only barcodes with a high probability of tagging a cell-containing droplet (i.e. we kept cells with an EmptyDrops FDR < 0.0001)(*116*). We removed cells labeled as doublets or ambiguous by demuxlet, keeping only barcodes labeled as singlets. We also filtered the data to include only high quality cells expressing between 3-20% mitochondrial reads and expressing more than 1,500 genes. We normalized data from each 10X sequencing lane using SCTransform in Seurat (*117*, *118*). In total, we obtained 42,488 high quality cells. We then merged data from each of the 10X lanes from all replicates, scaled the data, and ran principal components analysis using 5,000 variable features. We then integrated data with Harmony to correct the PCA embeddings for batch effects and individual effects (*95*).

Clustering and cell type annotation

To cluster the data, we applied Seurat's FindNeighbors using 100 dimensions from the Harmony-corrected reduced dimensions, followed by FindClusters at resolutions 0.1, 0.5, 0.8, and 1.

We performed differential expression analysis using the *limma* R package (*119*). First, we filtered genes to include only those expressed in at least 20% of cells in at least one cluster at a given clustering resolution. We then calculated pseudobulk expression values for each individual-replicate-cluster grouping (i.e. cells from the same individual, same replicate, and same cluster assignment). Accordingly, we define pseudobulk expression values as the sum of normalized counts in each group. Next we TMM-normalized pseudobulk expression values and used *voom* to calculate a weighted gene expression value to account for the mean-variance relationship (*120*). We then fit the following linear model:

$$Y = 0 + \beta\_cluster * x + \beta\_replicate * x + \beta\_individual * x$$

We used contrasts to first test for differential expression of each cluster compared to all other clusters and then to test for differential expression between pairs of similar clusters to find distinguishing markers. We annotated cell type identity of each cluster based on significant differential expression of the known marker genes.

We next compared cells to reference data sets of primary fetal cell types, Day 20 hESC-derived EBs, and hESCs (*99*, *100*). To integrate our cells with the reference sets, we first subset each reference set to include only protein coding genes. Because the Cao et al. reference set is so large, containing over 4 million cells, we subset cells from this reference set to include a maximum of 500 cells per cell type. We then normalized each reference set using SCTransform (*117*, *118*). We then merged the data sets using Seurat, re-ran SCTransform regressing out data set specific effects of sequencing depth, scaled the data, and ran principal components analysis. We then ran *Harmony* to correct PCA embeddings for the effects of each data set to complete the integration(*95*). We then transferred cell type annotations from cell types present in the fetal

reference and hESC to EB cells. For each EB cell, we found the 5 nearest reference cell in Harmony-corrected PCA space based on Euclidean distance. If 3/5 of the nearest reference cells shared the same annotation, we transferred the cell type annotation to the EB cell. If fewer than 3 of the nearest reference cells shared their annotation, we label the EB cell as 'uncertain.

We also performed topic modeling using *FastTopics* to learn major patterns in gene expression within the data set, or topics, and model each cell as a combination of these topics(*103*, *104*). For this analysis, we used raw counts and filtered to include gene expressed in at least 10 cells. We then pre-fit a Poisson non-negative matrix factorization by running 1,000 EM updates without extrapolation to identify a good initialization at values of k equal to 6, 10, 15, 25, and 30. We used this initialization to fit a non-negative matrix factorization using 500 updates of the scd algorithm with extrapolation to identify 6, 10, 15, 25, and 30 topics. We then used *FastTopics*' diff_count_analysis function to identify genes differentially expressed between topics (*103*, *104*). We used these differentially expressed genes to interpret the cellular functions and identities captured by each topic. In some cases, differentially expressed genes included known marker genes (Table 3-1).

Hierarchical clustering based on cell type composition and gene expression

To understand how similar cell type composition is between replicates and individuals, we first calculated the proportion of cells from each individual in each replicate assigned to each Seurat cluster at resolution 0.1. Then, using the base R *heatmap* function, we visualized the clustering of these replicate-individual groups. This function performs hierarchical clustering on samples using the complete linkage method. We repeated this analysis using Seurat clusters at resolution 0.5, 0.8, and 1 to show that the overall patterns of hierarchical clustering are robust to cluster resolution. We performed an analogous analysis using topic loadings instead of cluster

proportions. Here, we determined the loading of each topic on cells from the same individual and batch, then used the same hierarchical clustering in heatmap to visualize patterns of similarity between cells of each individual and batch.

We also performed hierarchical clustering on gene expression of individual cells. To do so, we took the pseudobulk expression for each individual-replicate-cluster group and filtered to genes expressed in at least 20% of cells in at least one cluster. We then calculated the log10(counts per million) expression of each gene. We then generated a heatmap using the *ComplexHeatmap* R package, performing hierarchical clustering based on the complete linkage method (*121*).

Variance partitioning

Using the same pseudobulk data and precision weights computed by voom from differential expression analysis, we used the VariancePartition R package to quantify the variation attributable to cluster, replicate, and individual (*106*). We fit a random effect model and modelled cluster, replicate, and individual as random effects. We performed this analysis across all tested Seurat clustering resolutions (0.1, 0.5, 0.8, 1). We performed this analysis using both pseudobulk samples of cells from the same cluster, replicate, and individual and at single cell resolution with each cell as a sample. We also partitioned the variance in each Seurat cluster separately using a random effect model with terms for replicate and individual. For this analysis, we used pseudobulk samples of cells from each individual and replicate.

Trajectory inference and identification of dynamic gene modules

We inferred trajectories using PAGA in Scanpy using Seurat clusters at all tested resolutions (*107, 122, 123*). We assigned pseudotime using diffusion pseudotime with the

pluripotent cells assigned as the root (*109*). We then manually traced known developmental trajectories supported by the coarse-grained PAGA graph. At clustering resolution 1, we traced the trajectory from pluripotent cells to endothelial cells, hepatocytes, and neurons.

We then isolated cells from each of these three trajectories and use Split-GPM to simultaneously cluster samples and identify dynamic gene modules (*92*). For this analysis, we divided data into decile pseudotime bins and calculated pseudobulk gene expression for cells of the same individual, replicate, and pseudotime bin. We identified 5 dynamic gene modules in each trajectory and interpreted them using gene set enrichment. To understand the variation in dynamic gene expression between individuals and replicates, we re-ran split-GPM ten times and observed how often cells from each individual and replicate were assigned to the same sample cluster.

Supplementary Figures and Tables

Supplementary figures and tables for this chapter are included in Appendix A: Supplementary Figures and Tables.

Chapter IV: A comparative analysis of gene expression in embryoid body-derived cell types from humans and chimpanzees

Abstract

Phenotypic differences between humans and our closest living evolutionary relatives, chimpanzees, are hypothesized to be caused primarily by differences in gene regulation. In particular, gene regulatory changes occurring during development may underlie morphological differences between species. However, comparative genomic studies have been limited due to the inaccessibility of relevant biological samples. In this study, we differentiate embryoid bodies (EBs) from a panel of human and chimp induced pluripotent stem cells. EBs contain diverse cell types and enable access to mature and developmental cell types which are difficult or impossible to collect from primary tissues. We leverage this *in vitro* differentiation to compare human and chimpanzee gene regulation, identifying conserved and derived patterns of gene expression across diverse cell types. We determine that 92% of tested genes are differentially expressed between species in at least one cell type. We also observed a set of conserved genes which are never differentially expressed in any cell type, and which show enrichment for processes related to cellular maintenance.

Introduction

Gene regulatory differences are hypothesized to underlie morphological and behavioral differences between humans and our closest living evolutionary relatives—chimpanzees (*63*). To explore the molecular changes that have produced human-specific traits, many studies have compared gene regulatory phenotypes between primates in a variety of tissue and cell types (*19*, *23*, *51–53*, *55*, *56*, *58–60*, *124*). These studies have characterized evolutionary patterns of gene expression levels, DNA methylation levels, chromatin modifications and accessibility, protein expression levels, and 3D genome structure across many tissues. These studies have improved our understanding of gene regulatory mechanisms, identified regulatory processes under selection, and revealed the genetic underpinnings of some human adaptations.

Comparative studies using primate tissues have, however, suffered from technical challenges. Before 2015, primary primate tissues were collected opportunistically and without optimal, balanced study design(*125*). This has made it difficult, or impossible, to account for the contribution of environmental and technical variation to observed differences in gene regulation. Moreover, it has been difficult to obtain multiple tissues from the same individual. So, most studies have been limited to within-tissue comparisons because confounding tissue and individual precludes the identification of tissue-specific patterns of variation between species. In 2015, new regulations made it illegal to collect new samples from chimpanzees(*47*, *48*); thus, all comparative genomic studies performed since then have, and will continue to, rely on frozen tissue or self-renewing cell lines collected before 2015.

Many of the challenges associated comparative genomic studies in humans and chimpanzees can now be overcome with use of induced pluripotent stem cells (iPSCs) (*126*).

iPSCs are self-renewing and can be differentiated *in vitro* to produce many cell types. iPSC lines have been generated from many human individuals as well as a small panel of chimpanzee individuals, enabling comparisons of gene regulation with balanced studies, in controlled culture conditions. To date, the Gilad lab and others have compared gene expression, DNA methylation, and chromatin conformation between humans and chimpanzees in iPSCs (*54*, *61*, *127*). iPSC-derived cell types have further been leveraged for comparative studies of gene expression dynamics during two disease-relevant processes: response to hypoxic stress and cellular differentiation (*23*, *128*). Comparative studies of differentiation, in particular, have yielded insights into species-specific regulatory patterns which may be associated with craniofacial morphology and with proliferation of neuronal cell types (*129*, *130*). Such studies have also highlighted the conserved patterns of regulation between species, especially early in development. To continue these efforts to elucidate the gene regulatory mechanisms that separate humans and chimps, it will be useful to study many more cell types. Directed differentiation protocols can be used to produce many as-yet-unstudied cell types, but these protocols are often laborious, costly, and time-consuming. Recently, single cell sequencing has unlocked the potential of organoids and embryoid bodies to enable efficient and cost-effective analysis of many cell types in a single study.

Embryoid bodies (EBs) are three dimensional aggregates of iPSCs spontaneously differentiating to cells of all 3 germ layers (*45*). EBs do not produce pure cell populations, like directed differentiations, that can be meaningfully analyzed with bulk RNA-seq data; instead, EBs are composed of highly heterogeneous cell types representing diverse functionalities across many stages of differentiation. Using single cell RNA-seq, EBs can be computationally dissected and each component cell type can be analyzed.

In this study, we generated EBs from human and chimpanzee iPSCs to simultaneously study a multitude of cell types, including early developmental cell types and collected single cell RNA-seq data. We identify both conserved and species-specific patterns of gene expression across diverse cell types. The cell types derived from chimp embryoid bodies represent a new resource for comparative studies of early development.

Results

Study design, data collection, and preprocessing

To perform a comparative study of embryoid body differentiation, we used a panel of 6 human iPSC lines (3 LC-derived YRI lines and 3 fibroblast-derive lines) and 3 chimpanzee iPSC lines (*131–133*). For human YRI and chimpanzee lines, we generated EBs in 3 replicates. We generated only one replicate of the fibroblast-derived human iPSC lines. We collected scRNA-seq using the 10x platform after 21 days of differentiation. Each collection batch corresponded to a replicate of EB generation and included all human and chimp lines. After filtering, and quality control (Methods), we retained a sample of 69,469 high quality cells from humans and 45,741 high quality cells from chimpanzees (Fig. S4-1). We then normalized counts in each species separately, and integrated cells using *Harmony*, which anchors the data sets by cell type (*95*).

Clustering and Cell type composition

To identify cell types in the integrated set of human and chimpanzee EB cells, we ran uniform manifold approximation (UMAP) and identified clusters using the Louvain algorithm (*97*, *98*). Because we did not have an expectation for the number of discrete cell types we would find in this data, we clustered at multiple resolutions identifying 13 clusters at resolution 0.1, 26 clusters at resolution 0.5, and 35 clusters at resolution 1 (Fig. S4-3). We then visualized the expression of known marker genes in UMAP space (Fig. S4-2), and identified 8 broad cell type

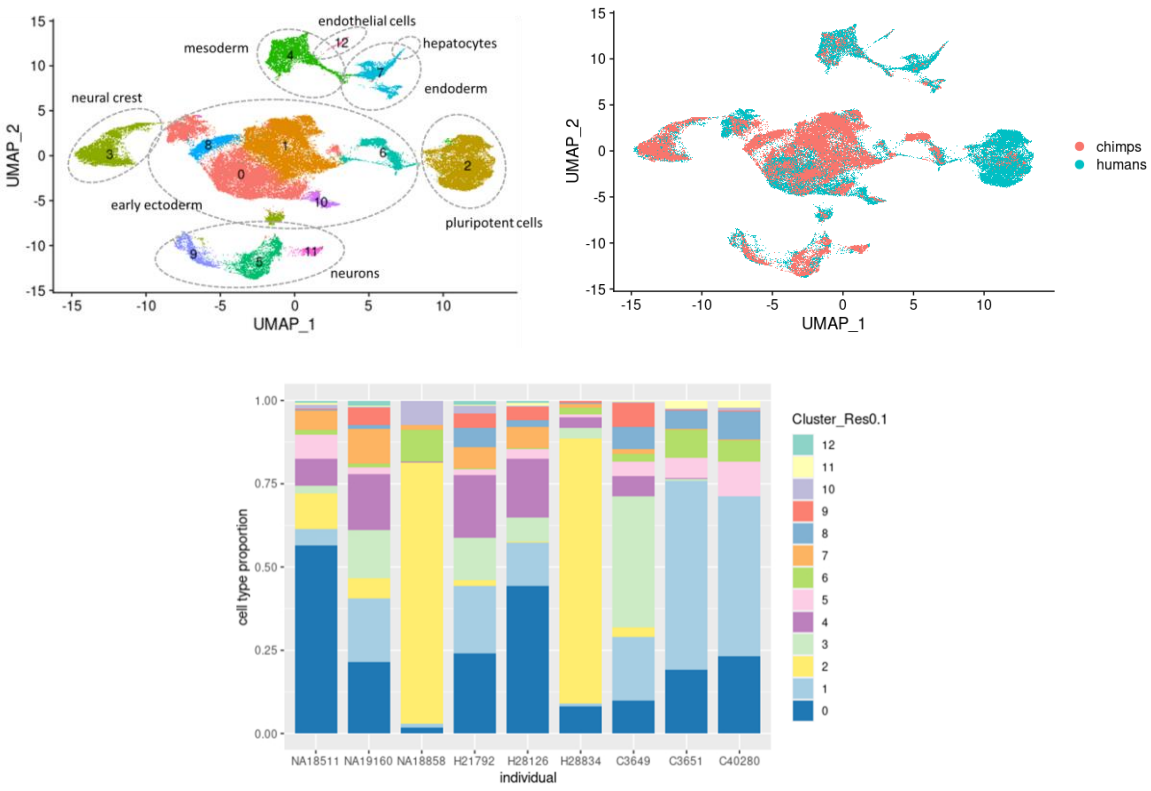


Fig. 4-1. Cell type composition of human and chimp EB cells. Top left) Cells are colored by cluster identity at clustering resolution 0.1. Broad cell type categories, determined by marker gene expression, are circled. Top right) Cells are colored by species. Bottom) Proportion cells assigned to each cell type (clusters from clustering resolution 0.1) in each individual. Individuals NA18511, NA19160, NA18858, H21792, H28126, and H28834 are humans. Individuals C3649, C3651, and C40280 are chimpanzees.

categories—pluripotent cells, endoderm, mesoderm, early ectoderm, neural crest, neurons, endothelial cells, and hepatocytes (Fig. 4-1). Clustering, at all tested resolutions, identifies subclusters within these broad cell type categories (Fig. S4-3). In the current analysis, we do not attempt to annotate these subclusters in more detail, simply using cluster definitions at each resolution as a proxy for cell type.

We observed that EBs from both species produce cells from all three germ layers, and cells from each species are represented in each cluster at clustering resolution 0.1 (Fig. 4-1, Fig. S4-2), indicating that our EB differentiation protocol produces similar cell types in both humans

and chimps. The proportion of cells in each cluster varies significantly between individuals (Fig. 4-1). Two human individuals—18858 and H28834—had a high proportion of cells assigned to cluster 2 at resolution 0.1. Because cluster 2 expressed pluripotency markers, we concluded that many of the cells from these two lines failed to differentiate (Fig. 4-1). To characterize patterns of variation in cell type composition, we performed hierarchical clustering to compare the proportions of cell in each cluster (at clustering resolution 0.1) between cells from each individual in each batch (Fig.S4-4) . We observed 18858 and H28834 do cluster separately from all other lines due to poor differentiation efficiency and increased representation of pluripotent cells. Among all other lines, samples tend to cluster by species and by individual rather than clustering by batch. While cell type proportions vary between individuals and between species, cell type proportions are consistent between replicates of the same line.

Characterizing the effects of biological and technical factors on gene expression variation

We first explored broad patterns of variation in gene expression. Using only the YRI human lines and chimpanzee lines (the lines for which we had multiple replicates), we performed variance partitioning using pseudo-bulk expression level of cells from the same cluster, replicate, and individual to estimate the relative contribution of cluster (cell type), species, individual, replicate, and sex to overall patterns of expression variation[(Fig. S4-5)(*106*). Using clusters defined at resolution 0.1, we found that cell type identity contributes by far the most variation, explaining a median value of ~45% of variation). Each of the remaining variables explains a median value of less than 10% of the total variation, with species explaining slightly more than individual, batch, or sex. A median value of ~25% of variance is explained by residuals, which can be attributed to noise or other technical variation not captured in the model. These patterns

hold true across all clustering resolutions, with cell type identity explaining the majority of variation, followed by species, individual, batch, and sex (Fig. S4-5).

Comparative assessment of gene expression across cell types

Next, we sought to identify conserved patterns of gene expression as well as tissue-specific differences in gene expression between species. To do so, we calculated pseudo-bulk gene expression levels from cells of the same cluster, individual, and replicate. We then tested for differential expression between species separately in each cluster, across all resolutions, using *dream* and *voom* (*120*, *134*). To improve power to detect differential expression in each cluster, and to find shared and cell-type specific differentially expressed genes, we used *mash* (*135*).

Overall, we found that 91.9% of tested genes are significantly differentially expressed between species in at least one cell type, at any clustering resolution. Conversely, 1341 genes (8% of all tested genes) were *not* significantly differentially expressed between species in any cell type, at any resolution, possibly indicating regulation of these genes has been evolutionarily conserved. We observed, however, that many of these never-differentially-expressed genes are expressed at low levels compared to genes that were found to be differentially expressed in at least one cluster (Fig. S4-6). We may simply lack the power to identify differential expression of lowly expressed genes and the regulation of these genes may not truly be conserved. For this reason, we subset the group of never-differentially-expressed genes to only the top 200 mostly highly expressed genes. We then perform gene ontology analysis to characterize the functions of these 200 genes with conserved patterns of regulation (Table S4-4)(*136–138*). Many of the enriched GO terms are related to core cellular processes and maintenance, including “mRNA processing” and “mRNA metabolic processes”). We also see strong enrichment for terms related to regulatory processes, particularly protein and DNA modifications.

To identify shared patterns of inter-specific expression differences between cell types, we computed the proportion of significantly differentially expressed genes shared by each pair of clusters at clustering resolution 0.1 (Fig. 4-2). After performing hierarchical clustering of cell types based on these proportions of shared effects, we see that cell type clusters 0,1,8, and 10, which all represent early ectodermal cell types, show similar patterns of differential expression. Similarly, we observe that cell type clusters 4,7, and 12 cluster together, because these are all non-pluripotent, non-ectodermal cell types (Fig. 4-1). Cluster 6, which likely represents early developmental cells transitioning from a pluripotent state into an ectodermal state, has the most distinct pattern of expression differences between species. As expected, these results show that closely related cell types share differences in gene regulation between species.

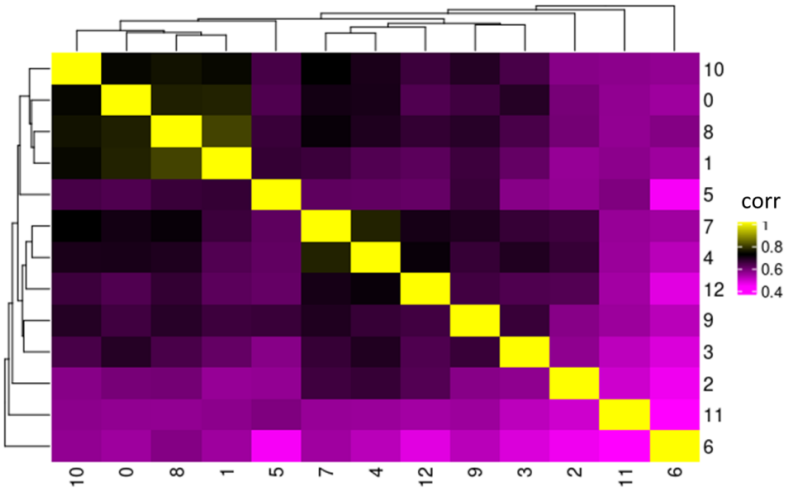


Fig. 4-2. Shared patterns of DE between cell types. Heatmap showing hierarchical clustering of cell type clusters (identified at clustering resolution 0.1) based on the correlation of effect sizes from differential

We also identified cell type-specific patterns of differential expression between humans and chimpanzees at clustering resolution 0.1 (Table 4-1). These cell-type specific differences may underlie phenotypic differences between species. To characterize the potential functional implications of these differences, we perform gene ontology analysis (Table S4-5. Table S4-5, Table S4-7, Table S4-8) (*136–138*). Here, I show the results from several clusters (from

clustering resolution 0.1) which may reveal interesting divergence in gene regulation between species, but which warrant further investigation.

In cluster 4 (representing mesodermal cells), cell type-specific DE genes are enriched for many gene sets involved in cellular responses to particular signaling events and environmental exposures (ex: response follicle stimulating hormone, response to gonadotropin, response to

Cluster Number (Res 0.1)	# of DE genes	# of cluster-specific DE genes
0	8394	42
1	8501	62
2	5528	276
3	7154	204
4	5911	37
5	7180	166
6	8458	953
7	5721	20
8	6690	25
9	6320	115
10	6002	39
11	6765	756
12	4441	156

Table 4-1. Number of differentially expressed genes in each cell type. The number of significantly differentially expressed genes identified in each cluster (at clustering resolution 0.1) and the number of cell type-specific DE genes identified in each cluster.

cadmium ion, etc.) (Table S4-5). These gene set represent regulatory pathways activated only in specific contexts, potentially making then less evolutionarily constrained (*19*). Further investigation into the genes driving these enrichments is necessary to understand whether differences in regulation of these genes could explain phenotypic differences between species.

We also observe that in several mature cell type clusters present in the data, cell-type specific differentially expressed genes show enrichment for GO terms related to developmental processes cell type specific functions (*136–138*). For example, dell type-specific DE genes identified in cluster 5 (representing neuronal cells) are enriched for the GO term "neuronal development", as well as terms related to cilia assembly. Non-motile "primary" cilia are known to play a role in neuronal cell fate and differentiation (Table S4-6)(*139*). And, cell type-specific DE genes from cluster 7 (endodermal cells) are enriched for genes related to fate commitment (Table S4-8). These results are somewhat unexpected; it is unlikely, that core genes related involved in to the development of particular cell types have diverged between humans and chimps. Instead, these observed patterns of cell type-specific DE could be explained by slight differences in the maturity of cell types produced by this differentiation protocol in each species. For example, chimp neuronal cells may achieve a slightly more mature state than human neurons in these differentiation conditions, leading to differential expression of genes related to neural development. In these examples, clustering of cells at higher resolution may better account for these differences in cell type maturity, but could also result in lower power to detect differentially expressed genes due to lower numbers of cells, and possibly fewer individuals, in each cluster. In the discussion below, I propose potential methods to account for species-biased cellular heterogeneity.

Next, we were interested in characterizing the functions of cell type-specific DE genes identified in early developmental cell types because comparative studies of these cells has never before been possible. In particular, we were interested in genes specifically DE in cluster 6—the cluster of cells transitioning from pluripotency to early ectodermal fate. Using gene ontology analysis, we observed that genes specifically differentially expressed in cluster 6 are enriched for several GO terms related to protein localization (Table S4-7)(*136*–*138*). We also observed that cluster 6-specific DE genes are enriched for genes associated with stem cell population maintenance, which may, again be indicative of species-biased differences in cellular maturity. These results indicate that while embryoid bodies from humans and chimps hold great promise as a model system for comparative studies of early development, there is additional work to be done to gain biological insight into diverged patterns of gene regulation.

Discussion

By generating a diverse set of cell types from humans and chimpanzees using iPSC-derived embryoid bodies, we were able to compare gene expression patterns between species in a breadth of cell types, including developmental cell types. We found that the vast majority of genes are significantly differentially expressed between species in at least one cell type, in at least one condition. We also observed shared patterns of inter-species differential expression between closely related cell types. Additionally, we observed that genes which are never significantly differentially expressed in any cell type, showing evidence for evolutionarily conserved regulation, are enriched for core processes involved in cellular upkeep.

We identified cell type-specific patterns of differential expression between species which require further investigation. In particular, we observed that sets of cell-type specific DE genes were often enriched for GO terms related to developmental processes. We hypothesize that these

instances of differential expression could be explained by species-bias in maturation of cells within a cell type. Several approaches could be used to capture and correct for this potential effect. As a first step, identifying cell types at higher clustering may help subdivide cell types to capture effects of cell type heterogeneity that differ between species, including heterogeneity due to temporal variation. A caveat, however, is that further subdividing cell type categories will reduce the number of cells per cluster and will reduce power to detect differential expression (Table S4-2). Alternatively, it may be beneficial to learn patterns of temporal variation within a cell type category and correct for this variable in the differential expression model. For example, one could assign pseudotime values to cells (either within a cluster, or across the whole data set) and correct this variable (*109*). This method should maintain power to detect differential expression by retaining a larger number of cells per cell type category.

Future analysis of this data set should also include more thorough cell type annotation. In chapter 3, I integrated a subset of these cells (the human YRI individual) with a reference data set containing 77 human cell types(*99*). I showed that the transcriptomic profiles of the YRI cells closely match the profiles of 68/77 cells types. We can infer that chimp EBs, which contain a similar array of cell types as human EBs, will also contain many cell types that can be a useful model of primary cell types. This analysis will have the added benefit (and caveat) of identifying many relatively homogenous cell types, but with many fewer cells. Again, this will limit the number of cell types that have sufficient representation of individuals from each species for differential expression analysis. Nonetheless, genes that are significantly differentially expressed in these reference-based cell type groups may provide more meaningful biological insight and a better framework for contextualizing results than cell types learned from clustering and marker gene expression.

A key outcome of this work will the generation of a developmental chimpanzee atlas—the first resource of its kind. Furthermore, this work establishes that like human EBs, chimpanzee EBs are a useful model for genomic studies. As an *in vitro* system, EBs from both species can be used in further studies which can assay and compare a multitude of regulatory phenotypes to continue to deepen our understanding of the gene regulatory differences that underlie human-specific traits.

Methods

Samples

We used iPSC lines from 6 human individuals and 3 chimpanzees to form EBs (*131–133*). This panel includes 3 YRI iPSC lines that were reprogrammed from lymphoblastoid cell lines (*131*). The other 3 humans and all 3 chimps are from a panel of iPSCs reprogrammed from skin fibroblasts(*132, 133*). All lines have been previously characterized. Information about the sex of each of these lines can be found in Table S4-3.

iPSC maintenance

We maintained feeder-free iPSC cultures on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with StemFlex Medium (A3349401, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). We grew cells in an incubator at 37°C, 5% CO2, and atmospheric O_2. Every 3-5 days thereafter, we passaged cells to a new dish using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded cells with ROCK inhibitor Y-27632 (ab120129, Abcam).

Embryoid body formation and maintenance

We formed EBs using a modified version of the STEMCELL Aggrewell400 protocol. Briefly, we coated wells of an Aggrewell 400 24-well plate (34415, STEMCELL) with anti-adherence rinsing solution (07010, STEMCELL). We dissociated iPSCs and seeded them into the Aggrewell400 24-well plate at a density of 1,000 cells per microwell ($1.2x10^6$ cells per well) in Aggrewell EB Formation Medium (05893, STEMCELL). After 24 hours, we replaced half of the spent media with fresh Aggrewell EB Formation Medium. 48 hours after seeding the Aggrewell plate, we harvested EBs and moved them to an ultra-low attachment 6-well plate (CLS3471-24EA, Sigma) in E6 media (A1516401, ThermoFisher Scientific). We maintained EBs in culture for an additional 19 days, replacing media with fresh E6 every 48 hours. Three independent replicates of EB differentiation were performed from each of the three human YRI lines and for each of the chimpanzee lines. A single replicate of EB differentiation was performed for the 3 fibroblast-derived human lines.

Embryoid body dissociation

We collected and dissociated EBs 21 days after formation. We dissociated EBs by washing them with phosphate-buffered saline (PBS) (Corning 21-040-CV), treating them with AccuMax (STEMCELL 7921) and incubating them at 37°C in for 15-35 minutes. After 10 minutes in Accumax, we pipetted EBs up and down with a clipped p1000 pipette tip for 30 seconds. We repeated pipetting every five minutes until EBs were completely dissociated. We then stopped dissociation by adding E6 media and straining cells through a 40um strainer (Fisherbrand 22-363-547). We resuspended cells in PBS and counted them with a TC20 Automated Cell Counter (450102, BioRad). Before scRNA-seq, we mixed together an equal number of cells from each line.

Single cell sequencing

We collected scRNA-seq data using the 10X Genomics V3.0 kit. From the first replicate of EB differentiations, we mixed even numbers of cells from all 9 individuals, cells were collected across 9 lanes of a 10x chip, targeting 10,000 cells per lane. In this replicate, reagents from 3 different 10x kits were used. From replicates 2 and 3 of EB differentiation, EBs were only generated from the YRI individuals and the chimpanzees (6 individuals total). In each replicate, we mixed even numbers of cells of each individual and collected cells in 4 lanes of a 10x chip, targeting 10,000 cells per lane, and samples were processed using reagents from a single 10x kit.

Libraries were sequenced using paired-end 100 base pair sequencing on the HiSeq 4000 in the University of Chicago Functional Genomics Core. For libraries from replicate 1, we mixed equal proportions of each of the 6 out of the 9 libraries and sequenced the pooled samples on 1 lane of the HiSeq 4000. Preliminary analyses showed that 2 of these lines were low quality. We remade one of the low quality libraries and discarded the other. We then mixed equal proportions of the remade library with the remaining 3 libraries from replicate 1 and sequenced the pooled samples on one lane of the HiSeq 4000. Preliminary analyses indicated that 3 out of 4 of these libraries were below optimal quality, but would produce usable data. We then pooled together samples from the final 8 libraries from replicate 1, mixing equal parts of each of the 5 high quality libraries with half the amount of the other 3, and deep-sequenced this pool on 8 lanes of the HiSeq 4000. For replicate 2 libraries, we mixed equal parts of all 4 libraries and sequenced on 1 lane. After confirming that each library was high quality, we deep-sequenced the same pool on 6 additional lanes of the HiSeq 4000. For replicate 3 libraries, we mixed equal

parts of all 4 libraries and sequenced on 1 lane. After confirming that each library was high quality, we deep-sequenced the same pool on 4 additional lanes of the HiSeq 4000. In all cases, the number of lanes for deep sequencing was calculated to reach 50 % saturation, based on CellRanger's estimate of saturation from the initial lane of sequencing(*140*).

Generation of a set of orthologous exons

In order to exclude annotation differences as a source of differential expression between humans and chimpanzees, we generated a set of orthologous exons. We followed the approach laid out in Pavlovic 2018 and Blekhman 2012 with minor modifications(*141*, *142*). We began with the Ensembl version 100 release of the homo sapiens hg38 genome annotation. We filtered this set of annotations for protein coding exons. To identify potentially orthologous exons in chimpanzee genome version panTro6, we ran pblat) on the starting set of 1,457,971 protein coding human exons from 19,970 genes(*143*). We filtered the resulting chimpanzee set for exons that 1) did not have insertions or deletions greater than 20% of exon length and 2) mapped with percent identity of at least 80%. To resolve cases where multiple chimpanzee mappings met these criteria for a single human exon, we calculated the number of chimpanzee exons within 100kb that were also within 100kb of the human exon in the human annotation. We selected the chimpanzee exon that had the most neighboring exons in common with the exon in the human annotation. This resulted in a set of 1,383,887 orthologous exons from 19,935 genes.

We filtered this set further with two more pblat alignments. First, we removed exons that did not return the original location in humans when running pblat on the chimpanzee exons aligned to the human genome. Next, we removed exons that did not return the same location in chimpanzee when running pblat on the chimpanzee exons aligned to the chimpanzee genome. Finally, we removed exons that resulted in transcripts that spanned multiple contigs.

This resulted in a final set of 1,356,853 orthologous exons from 19,860 genes, with 93% of the exons in the full human annotation represented.

Alignment, demultiplexing, and cell filtering

We aligned sequencing reads using 10x genomics' cellranger pipeline (*140*). We generated three references using the cellranger mkref command: a human reference using hg38 and the orthologous exon set described above, a chimpanzee reference using panTro6 and the orthologous exons, and a combined human chimp reference. We aligned the sequencing reads to all three references. We used the combined reference to identify the species of cell-containing droplets. We then proceeded with count matrices for these cells using the alignments to each species individually.

We demultiplexed the samples using demuxlet (*113*). We obtained vcf files for YRI lines from the 1000 genomes project. We obtained vcfs for the human and chimpanzee panel from Greg Wray's lab. One individual had not previously been genotyped. To generate a vcf for this individual we used the GATK best-practice pipeline on the 7 RNA-seq samples generated in Pavlovic 2018 (*127*). We combined the resulting vcfs for each species individually and filtered for variants that were within orthologous exons.

We filtered the resulting cells using the following criteria. We removed cells marked as doublets or ambiguous by demuxlet. We removed cells with fewer than 0.1% or more than 20% of reads mapping to mitochondrial genes. We removed cells with fewer than 1000 genes detected.

Normalization, Integration and clustering

We first divided the cells by species and normalized cells from humans and chimps separately using SCTransform in Seurat (regressing out the effects of replicate and 10x lane, and using 6000 variable features)(*118*). We identified the common variable features between humans and chimps (3996 genes in total) and ran principle components analysis using only these genes. We integrated the human and chimp data sets using Harmony to correct PCA embeddings for species (*95*). To identify cell types in the integrated data set, we ran UMAP dimensional reduction based on the *Harmony* reduction and clustered using the Louvain algorithm at resolutions 0.1, 0.5, and 1 (*97*, *98*).

To characterize broad categories of cell types present in the data, we visualized expression of a variety of known marker genes (Fig S4-2).

To understand how similar cell type composition is between replicates and individuals, we first calculated the proportion of cells from each individual in each replicate assigned to each Seurat cluster at resolution 0.1. We then performed hierarchical clustering using *ComplexHeatmap*, calculating the Pearson correlation of each pairwise comparison between samples, and clustering samples using the complete linkage method(*121*).

Variance Partitioning

For this analysis, we excluded the fibroblast-derived human lines. We calculated pseudobulk expression levels for cells of the same cluster (using clusters defined at resolution 0.1), individual, and batch. We normalized the pseudobulk counts using TMM and cyclic loess, then and used *voom* to calculate a weighted gene expression value to account for the mean-variance relationship (*120*). Using the *VariancePartition* R package, we fit a random effect

model and modelled cluster, species, individual, replicate, and sec effects to quantify the variation attributable to each of these factors(*106*).

Comparative assessment of gene expression across clusters

To test for differential expression between species, we subset cells from each cluster across all resolutions. For each cluster, we calculated pseudobulk expression levels for cells of the same individual and batch. We only keep pseudobulk samples representing at least 5 cells and we only analyze clusters that have pseudobulk samples from at least 2 individuals of each species. We the filtered genes to include only those expressed at greater than 1 count per million (cpm) in at least one species. We normalized pseudobulk counts using TMM and cyclic loess, then and used *voom* to calculate a weighted gene expression value to account for the mean-variance relationship (*120*). Using the *dream* R package, we implemented a linear model to test for differences in gene expression between species while correcting for the effects of cluster, individual, and replicate(*134*). Significance and sharing of differentially expressed genes was calculated using multivariate adaptive shrinkage implemented using the *mashr* R package (*135*).

Gene Ontology Analysis

We performed enrichment tests for gene ontology terms using the *topGO* R package(*136*). We used a Fisher's exact test to test for over-representation of GO terms within a group of genes of interest. We tested for enrichment of gene ontology terms, biological process, cellular component, and molecular function(*137*, *138*). We apply this method first to the top 200 most highly expressed genes that were not found to be significantly differentially expressed between species in any cluster at any resolution to explore functions of genes with a conserved pattern of regulation. Then, we tested cell-type specific genes which were only significant in one cell type cluster (at clustering resolution 0.1). In this analysis, we subset to include only genes

that were tested in all clusters. The enriched GO terms in these analyses may represent functions for which regulation has diverged between humans and chimps.

Supplementary Figures and Tables

Supplementary figures and tables for this chapter are included in Appendix A: Supplementary Figures and Tables.

Chapter V: Discussion

Together, these projects leverage the power of iPSCs and iPSC-derived cell types to explore developmental gene regulation. First, I identified dynamic eQTLs, or genetic variants associated with differences in the change in gene expression level through differentiation time; many of the identified dynamic eQTLs are located in cell-type-specific regulatory elements and have previously been associated with heart phenotypes and cardiovascular disease. Of particular note, we identified a subset of dynamic eQTLs—"middle" dynamic eQTLs—with a transient effect on gene expression levels in the intermediate stages of cardiomyocyte differentiation. Middle dynamic eQTLs could not have been identified as static eQTLs in either iPSCs or cardiomyocytes, highlighting the power of timecourse studies that capture multiple stages of differentiation. Middle dynamic eQTLs, despite having a fleeting effect, are associated with cardiovascular diseases and human traits, as well as other gene regulatory processes. For example, one middle dynamic eQTL was also found to be a trans eQTL in GTEx coronary artery, suggesting that dynamic eQTLs acting during development can have long-term effects on the gene regulatory landscape and on human disease risk.

Having confirmed the value of dynamic eQTLs identified in the differentiation cardiomyocytes, I next sought out to establish a system of the discovery of many more dynamic eQTLs acting during the differentiation of a multitude of iPSC-derived cell types using embryoid bodies. In a pilot study of human EBs, I characterized the cell types and states produced by EB differentiation, quantified the effects of biological and technical factors, and explored dynamic patterns of variation in inferred developmental trajectories. This study had two particularly

notable outcomes: first, by integrating EB cells with a reference set of cells from fetal primary cell types, I was able to annotate 68 different cell types within EBs. This demonstrated that EB differentiation produces many immature, developmental cell types, but also produces cell types that are a useful *in vitro* model of primary fetal cells. Second, the results of this pilot study informed the study design for a large scale study of human EBs. At the time of writing, 54 human YRI cells lines have been differentiated to EBs, each in two independent replicates. In total, this experiment is expected to result in a dataset of over 1 million cells and will be used to identify eQTLs and dynamic eQTLs.

Lastly, in Chapter 4, I performed a comparative study human and chimpanzee embryoid bodies. This work identified conserved patterns of gene expression regulation across developmental cell types as well as tissue-specific patterns of differential expression. This work also represents a major contribution to the field, in that the chimpanzee EB dataset will be a new resource for the study of developmental cell types which have never before been accessible, either from primary tissue or form cultured cells.

Differentiation propensity of iPSC lines

In the cardiomyocyte differentiation project (Chapter 2), we discovered that cell lines clustered into two distinct groups based on their patterns of dynamic gene expression. One of these groups showed a greater increase in expression over time of genes associated with myogenesis, while the other group showed increased expression of gene related to KRAS activation. We hypothesized that these cell line clusters could be explained by differences in cell type composition and, relatedly, differences in differentiation efficiency. But, we could not test these hypotheses using only the bulk data collected in that study.

Building on the results of the bulk RNA-seq experiment, Dr. Reem Elorbany, previously a graduate student in the Gilad lab, performed another timecourse study of cardiomyocyte differentiation, this time collecting single cell RNA-seq (data not yet published). The main motivation for this study was to increase power to identify eQTLs by isolating homogenous populations of cells. The single cell data set did successfully achieve that goal; it also provided retroactive insight into the cause of the two cell line clusters observed in the original study. In the single cell data, three distinct terminal cell types could be identified: cardiomyocytes (expressing TNNT2), connective tissue (expressing COL3A1), and hepatocytes (expressing APOA1 and AFP). Cell lines that had been assigned to cluster 2 in the bulk experiment (the cluster with increased expression of KRAS-related genes) showed decreased expression of all of these markers relative to cell lines from the other cluster (data not yet published). This observation suggests that cluster 2 cell line produced fewer terminally differentiated cells and may have had lower differentiation efficiency overall in both the original study and in the follow-up. Two of the human iPSC lines used in the human embryoid body pilot study (Chapter 3) were also used in the cardiomyocyte time courses. Of these two, 18858 was assigned to cell line cluster 2 (the cluster of cell lines with lower cardiomyocyte differentiation efficiency) while 18511 was assigned to cell line cluster 1 (the cluster of cell liens with higher cardiomyocyte differentiation efficiency). In the EB differentiation, 18858 also showed lower differentiation efficiency that 18511. While more evidence is needed to draw robust conclusions, these observations provide anecdotal evidence that cell lines may have an intrinsic differentiation propensity that broadly effects differentiation efficiency across all lineages. As discussed in Chapter 3, this differentiation propensity is likely not due to donor-specific genetic factors, but could instead be caused by other cell line intrinsic factors due to environmental, technical, or biological events

experienced by the cells (*105*). Depending on when the event occurred, the observed differentiation propensity of a certain iPSC line may be shared among all iPSCs descended from a particular reprogramming batch or even shared only among cells of a particular cryopreserved aliquot. Hence, it cannot be assumed that an iPSC line that exhibited low differentiation efficiency in one experiment will differentiate poorly in all future experiments.

What is a cell type? Shifting paradigms

In Chapter 3, I used three separate approaches to annotate cells from human embryoid bodies. Specifically, I first performed unsupervised clustering implemented in Seurat, ran differential expression analysis to identify genes significantly upregulated in each cluster, and then assigned cell type labels based on upregulation of known marker genes. Next, I integrated our EB cells with reference data sets and annotated cells based on the overall similarity of their gene expression profile to reference cells. Lastly, I used topic modelling to learn major patterns in gene expression within the data set, or topics, and model each cell as a combination of these topics. There are advantages and disadvantages to each of these approaches, some practical and some conceptual. To contextualize these pros and cons, it is first necessary to understand that each approach is rooted in a distinct paradigm of how we conceptualize and define cell identity.

So, what *is* a cell type? The cells that make up the human body vary in many dimensions, exhibiting heterogeneous morphology, physiology (*144*, *145*), functionality, regulation and localization (in both space and time). Any of these features can be used to categorize cells, provided that they can be effectively measured. For this reason, cell type definitions have changed through time as technology has advanced. For example, when cells were first discovered, they could only be classified based on visible morphological differences under a microscope. Since then, cell type definitions evolved to incorporate measurements from

emerging technologies. Eventually, with the advent of micro-arrays and bulk RNA-seq, cell types came to be defined as discrete categories distinguished by expression of marker genes. Now, there is a wealth of literature cataloguing the biology of discrete cell types, their functions, and their involvement in human disease.

However, in the light of single cell RNA-seq, previously defined cell types have been broken down. scRNA-seq enables measurement of gene expression levels at maximum resolution, and produces high dimensional data sets measuring thousands of genes from an individual cell. A single cell, then, will fall within a continuous spectrum of variation. Indeed, single cell RNA-seq studies have characterized heterogeneity in "pure" cell type populations, identifying cell "subtypes". Realistically, scRNA-seq data sets can be infinitly subdivided, partitioning cells into smaller and smaller clusters (*145*). These subdivisions may even be useful, and could yield biological insight into the gene regulatory differences between subsets of cells. They are, nonetheless, artificial categories. And, ultimately, discrete categories fail to capture the full complexity of cellular heterogeneity.

In an alternate paradigm, we can adopt a continuous view of cell type identity, where a cell is defined by its location in high dimensional space--by its unique combination of continuous traits. This cell type definition will more closely approximate biological truth, with the limit of our understanding still relying on available measurement tools. However, this paradigm makes it difficult to interpret heterogeneity. If a cell is truly unique, it becomes impossible to compare and cannot be contextualized using the wealth of biological information gathered about discrete cell types. To make this paradigm more practical, methods like grades of membership modelling and topic modelling can be applied to single cells data (*103*). These methods define each cell as a combination of discrete groups, thereby providing a more continuous view of cell type identity

while still leveraging the practicality of discrete groups for comparison and interpretation. These discrete groups, however, typically provide a gene-centric view of the data, capturing modules of co-regulated genes (like the topics identified via topic modelling in Chapter 3). These gene modules can be interpreted based on differential expression of well-characterized genes or with gene set enrichment, but, because there is not a wealth of information characterizing these gene modules, they remain difficult to contextualize. To illustrate this point, compare the conclusion we were able to draw from topic modelling in EB cells and the annotation using a reference, both detailed in chapter 3. After reference integration, EB cells could be assigned to discrete cell type identities, immediately providing a labels like “cardiomyocyte” or “excitatory neuron” that immediately contextualize that cell in space and time, and provide intuitive insight into that cell’s function based on everything we, as a field, have learned about those “cell types”. From the topic modelling analysis, we labelled as cell as a combination of topics, and in some cases, were able to guess at the biological function of a particular topic based on upregulation of marker genes. This was suboptimal because these were marker genes of known discrete cell types, so interpretation of those topics was still restricted to the discrete paradigm.

In the end, we, as scientists, choose the definition of a cell type. Since the concept of a cell type is a convention, we should choose the definition pragmatically, to maximize biological insight (*146*). Cell type definition can be, and should be, flexible, bending to suit a particular hypothesis or experiment. In the short term, this may mean that the discrete cell type paradigm continues to dominate because it facilitates interpretation. However, in the long term, I expect the continuous paradigm, or at least, the semi-continuous paradigm enabled by topic modelling, to yield the most novel and nuanced insight because it more closely approximates biological truth.

Future directions for genomic studies of embryoid bodies

In Chapter 3, I showed that EBs have the potential to be a powerful model for the study of human gene regulation. First, EBs are composed of functionally and temporally diverse cells, some of which have highly similar transcriptome profiles to primary fetal cells. And, as an *in vitro*, iPSC-derived system, they enable efficient, cost effective, and high throughput studies of inter-individual variation in gene regulation. I foresee several key areas where EBs will facilitate important new discoveries. First, large-scale studies of EBs can be used to identify static eQTLs in developmental cell types, cell type composition QTLs, and dynamic eQTLs across diverse developmental trajectories. These analyses will likely reveal the mechanisms of many unexplained GWAS variants and will represent a major contribution to our understanding of genetic underpinnings of human disease. As other types of single cell assays mature, EBs will continue to be an important model system for comparison of gene regulation across cell types; soon, it will be possible to collect multi-model data capturing gene expression (scRNA-seq), chromatin accessibility (scATAC-seq), and protein levels (CITE-seq, Cell Hashing, REAP-seq) across all EB cell types (*147–149*). These datasets will enable studies of cell type specific associations between each of these regulatory phenotypes, providing a new window into patterns and mechanisms of gene regulation.

I also anticipate that studies of spatial transcriptomics in EBs will empower researchers to explore an array of novel questions. Because EBs contain temporally diverse cells, spatial transcriptomics can assign a cell to a specific space and time, together providing detailed information about context specific gene regulation. For example, spatial analysis of EBs will enable studies of the association of paracrine signaling and gene regulation. A variety of methods exist to capture the spatial localization of cells in three dimensional tissues and organoids. Using

these technologies, it is possible to characterize the regulatory profiles of a cell and its immediate neighbors. One could then broadly test for the association between neighboring cell states to understand the effects of paracrine signaling. This strategy could provide insight into the influence of paracrine signaling on cell fate specification, cellular differentiation and maturation, as well as rare context-specific regulatory events.

EBs will also facilitate the effects of environmental conditions across diverse cell types. With clever study design, Findley *et al.* efficiently tested the effects of many treatments conditions, across several cell types, and across several individuals, in a single experiment (*150*). While the sample size of this study was too small to have sufficient power to map eQTLs, they did identify significant allele specific expression (ASE) and condition-specific ASE (cASE). This analysis identified treatment effects on cis-regulation of gene expression, which are examples of gene by environment interaction. This study then showed that ~ 50% of genes with cASE were involved in complex traits based on GWAS, which is significantly greater than ASE or eQTL genes(*91*). This study demonstrated the importance of characterizing inter-individual variation in cell-type specific gene by environment interaction to understanding human traits and diseases. A similar approach could be taken with EBs to find inter-individual variation in treatment response across a much larger sample of cell types. With a large enough sample size, such a study could be used to find response QTLs acting in developmental cell types in many treatment conditions. There is a large body of evidence suggesting that environmental exposures during development have long-term health consequences, and people have even begun to use EBs as a model system to study the effects of nicotine exposure during development t(*44*, *151*). Future studies can expand the treatments tested to greatly improve our understanding of the interplay between environmental exposure, developmental gene regulation, and disease risk.

Concluding Remarks

Gene regulation is dynamic, changing dramatically during cellular differentiation. Genomic studies of have successfully characterized effects of non-coding variants across many human tissues and cell types, but have largely failed to capture the regulatory effects of genetic variants on gene expression dynamics during developmental processes. In this dissertation, I have begun to address this gap. Using a time course study of iPSC-derived cardiomyocyte differentiation, I identified dynamic eQTLs and demonstrated that these genetic loci, which are associated with change in gene expression level over time, are associated with long-term cardiovascular disease risk and heart phenotypes. I then explored the use iPSC-derived embryoid bodies as a model system for human genomic studies, demonstrating that they can be used to analyze inter-individual, and inter-species, variation in gene regulation across functionally and temporally diverse cell types and across a multitude of developmental trajectories. Overall this work underscored the important role of dynamic gene regulation in shaping human traits and disease, and provided foundational work to enable future genomic studies of regulatory dynamics in iPSC-derived cells.

References

1. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019), doi:10.1093/nar/gky1120.

2. D. Welter *et al.*, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* (2014), doi:10.1093/nar/gkt1229.

3. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* (2013), doi:10.1002/0471142905.hg0720s76.

4. M. Seifi, M. A. Walter, Accurate prediction of functional, structural, and stability changes in PITX2 mutations using in silico bioinformatics algorithms. *PLoS One* (2018), doi:10.1371/journal.pone.0195971.

5. A. A. Pai, J. K. Pritchard, Y. Gilad, The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genet.* (2015), doi:10.1371/journal.pgen.1004857.

6. T. A. Manolio *et al.*, Finding the missing heritability of complex diseases. *Nature*. **461**, 747–53 (2009).

7. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* (2012), doi:10.1126/science.1222794.

8. A. Battle *et al.*, Impact of Regulatory Variation from RNA to Protein. *Science (80-.).* **347**, 1922–2013 (2015).

9. N. E. Banovich *et al.*, Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genet.* **10** (2014), doi:10.1371/journal.pgen.1004663.

10. G. McVicker *et al.*, Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science (80-.).* **342**, 747–749 (2013).

11. P. Benaglio *et al.*, Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex trait variants using single nucleus ATAC-seq. *bioRxiv* (2020).

12. Y. Gilad, S. A. Rifkin, J. K. Pritchard, Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* (2008), , doi:10.1016/j.tig.2008.06.001.

13. F. Aguet *et al.*, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

14. D. W. Yao, L. J. O'Connor, A. L. Price, A. Gusev, Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* (2020), doi:10.1038/s41588-020-0625-2.

15. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* (2020), doi:10.1126/science.aaz1776.

16. B. D. Umans, A. Battle, Y. Gilad, Where Are the Disease-Associated eQTLs? *Trends Genet.* (2020), , doi:10.1016/j.tig.2020.08.009.

17. E. R. Gamazon *et al.*, Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* (2018), doi:10.1038/s41588-018-0154-4.

18. A. Gerrits *et al.*, Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet.* (2009), doi:10.1371/journal.pgen.1000692.

19. R. Blekhman, A. Oshlack, A. E. Chabot, G. K. Smyth, Y. Gilad, Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* (2008), doi:10.1371/journal.pgen.1000271.

20. D. A. Knowles *et al.*, Determining the genetic basis of anthracycline-cardiotoxicity by molecular response QTL mapping in induced cardiomyocytes. *Elife*. **7** (2018), doi:10.7554/eLife.33480.

21. L. B. Barreiro *et al.*, Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1204–9 (2012).

22. S. Kim-Hellmuth *et al.*, Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat. Commun.* (2017), doi:10.1038/s41467-017-00366-1.

23. M. C. Ward, N. E. Banovich, A. Sarkar, M. Stephens, Y. Gilad, Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocyteso Title. *bioRxiv* (2020) (available at https://www.biorxiv.org/content/10.1101/2020.03.28.012823v1.full.pdf).

24. K. Takahashi, S. Yamanaka, Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. **2**, 663–676 (2006).

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

26. J. Yu *et al.*, Induced pluripotent stem cell lines derived from human somatic cells. *Science (80-.).* (2007), doi:10.1126/science.1151526.

27. C. K. Burrows, N. E. Banovich, B. J. Pavlovic, K. Patterson, Genetic Variation , Not Cell Type of Origin , Underlies the Majority of Identifiable Regulatory Differences in iPSCs. *PLoS Genet.* **12**, 1–18 (2016).

28. P. W. Burridge *et al.*, Chemically defined generation of human cardiomyocytes. *Nat. Methods*. **11**, 855–860 (2014).

29. K. Si-Tayeb *et al.*, Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology* (2010), doi:10.1002/hep.23354.

30. S. M. Chambers *et al.*, Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nat. Biotechnol.* (2009), doi:10.1038/nbt.1529.

31. J. Zhang *et al.*, Functional Cardiomyocytes Derived from Human Induced Pluripotent Stem Cells. *Circ Res*. **104** (2010), doi:10.1161/CIRCRESAHA.108.192237.Functional.

32. N. E. Banovich *et al.*, Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 122–131 (2018).

33. P. Douvaras *et al.*, Directed Differentiation of Human Pluripotent Stem Cells to Microglia. *Stem Cell Reports* (2017), doi:10.1016/j.stemcr.2017.04.023.

34. S. Karumbayaram *et al.*, Directed differentiation of human-induced pluripotent stem cells generates active motor neurons. *Stem Cells* (2009), doi:10.1002/stem.31.

35. Y. Shi, P. Kirwan, F. J. Livesey, Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat. Protoc.* (2012), doi:10.1038/nprot.2012.116.

36. J. Muffat *et al.*, Efficient derivation of microglia-like cells from human pluripotent stem cells. *Nat. Med.* (2016), doi:10.1038/nm.4189.

37. Y. Li *et al.*, Induction of Expansion and Folding in Human Cerebral Organoids Article Induction of Expansion and Folding in Human Cerebral Organoids. *Stem Cell*, 1–12 (2017).

38. B. Song *et al.*, The Directed Differentiation of Human iPS Cells into Kidney Podocytes. *PLoS One* (2012), doi:10.1371/journal.pone.0046453.

39. L. Menendez *et al.*, Directed differentiation of human pluripotent cells to neural crest stem cells. *Nat. Protoc.* (2013), doi:10.1038/nprot.2012.156.

40. M. P. Walczak, A. M. Drozd, E. Stoczynska-Fidelus, P. Rieske, D. P. Grzela, Directed differentiation of human iPSC into insulin producing cells is improved by induced expression of PDX1 and NKX6.1 factors in IPC progenitors. *J. Transl. Med.* (2016), doi:10.1186/s12967-016-1097-0.

41. S. Kim-Hellmuth *et al.*, Cell type specific genetic regulation of gene expression across human tissues. *bioRxiv* (2019), , doi:10.1101/806117.

42. D. McDonald *et al.*, Defining the Teratoma as a Model for Multi-lineage Human Development. *Cell* (2020), doi:10.1016/j.cell.2020.10.018.

43. G. Rossi, A. Manfrin, M. P. Lutolf, Progress and potential in organoid research. *Nat. Rev. Genet.* (2018), , doi:10.1038/s41576-018-0051-9.

44. H. Guo *et al.*, Single-Cell RNA Sequencing of Human Embryonic Stem Cell Differentiation Delineates Adverse Effects of Nicotine on Embryonic Development. *Stem*

Cell Reports (2019), doi:10.1016/j.stemcr.2019.01.022.

45. J. Itskovitz-Eldor *et al.*, Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. *Mol. Med.* (2000), doi:10.1007/bf03401776.

46. H. Kurosawa, Methods for inducing embryoid body formation: in vitro differentiation system of embryonic stem cells. *J. Biosci. Bioeng.* (2007), doi:10.1263/jbb.103.389.

47. Fish and Wildlife Service, Endangered and Threatened Wildlife and Plants; Listing All Chimpanzees as Endangered Species. *Fed. Regist.* (2015), (available at https://www.federalregister.gov/documents/2015/06/16/2015-14232/endangered-and-threatened-wildlife-and-plants-listing-all-chimpanzees-as-endangered-species).

48. J. Kaiser, NIH to end all support for chimpanzee research. *Science (80-.).* (2015), (available at https://www.sciencemag.org/news/2015/11/nih-end-all-support-chimpanzee-research).

49. I. Friedrich Ben-Nun *et al.*, Induced pluripotent stem cells from highly endangered species. *Nat. Methods* (2011), doi:10.1038/nmeth.1706.

50. A. Birbair, Ed., *iPSCs from Diverse Species* (Academic Press, 2020).

51. W. Enard *et al.*, Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*. **418**, 869–72 (2002).

52. P. Khaitovich *et al.*, Evolution: Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science (80-.).* (2005), doi:10.1126/science.1108296.

53. C. E. Cain, R. Blekhman, J. C. Marioni, Y. Gilad, Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* (2011), doi:10.1534/genetics.110.126177.

54. L. E. Blake *et al.*, A comparison of gene expression and DNA methylation patterns across tissues and species. *Genome Res.* (2020), doi:10.1101/gr.254904.119.

55. M. Cáceres *et al.*, Elevated gene expression levels distinguish human from non-human primate brains. *Proc. Natl. Acad. Sci. U. S. A.* (2003), doi:10.1073/pnas.2135499100.

56. M. W. Karaman *et al.*, Comparative Analysis of Gene-Expression Patterns in Human and African Great Ape Cultured Fibroblasts, 1619–1630 (2003).

57. P. de Candia, R. Blekhman, A. E. Chabot, A. Oshlack, Y. Gilad, A Combination of Genomic Approaches Reveals the Role of FOXO1a in Regulating an Oxidative Stress Response Pathway. *PLoS One*. **3**, e1670 (2008).

58. C. C. Babbitt *et al.*, Both noncoding and protein-coding rnas contribute to gene expression evolution in the primate brain. *Genome Biol. Evol.* (2010), doi:10.1093/gbe/evq002.

59. Y. Shibata *et al.*, Extensive Evolutionary Changes in Regulatory Element Activity during

Human Origins Are Associated with Altered Gene Expression and Positive Selection. *PLoS Genet.* **8**, e1002789 (2012).

60. Z. Khan *et al.*, Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science (80-.).* (2013), doi:10.1126/science.1242379.

61. I. E. Eres, K. Luo, C. J. Hsiao, L. E. Blake, Y. Gilad, Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet.* (2019), doi:10.1371/journal.pgen.1008278.

62. F. C. Chen, W. H. Li, Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* (2001), doi:10.1086/318206.

63. M. C. King, A. C. Wilson, Evolution at two levels in humans and chimpanzees. *Science (80-.).* (1975), doi:10.1126/science.1090005.

64. S. H. Wang, C. J. Hsiao, Z. Khan, J. K. Pritchard, Post-translational buffering leads to convergent protein expression levels between primates. *Genome Biol.* (2018), doi:10.1186/s13059-018-1451-z.

65. Y. I. Li *et al.*, RNA splicing is a primary link between genetic variation and disease. *Science (80-.).* **352**, 600–604 (2016).

66. F. W. Albert, L. Kruglyak, The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

67. Z. Zhu *et al.*, Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48** (2016), doi:10.1038/ng.3538.

68. D. L. Nicolae *et al.*, Trait-Associated SNPs Are More Likely to Be eQTLs : Annotation to Enhance Discovery from GWAS. *PLoS Genet.* **6** (2010), doi:10.1371/journal.pgen.1000888.

69. R. Joehanes *et al.*, Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* (2017), doi:10.1186/s13059-016-1142-6.

70. X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* (2017), doi:10.1371/journal.pgen.1006646.

71. Gte. Consortium *et al.*, Genetic effects on gene expression across human tissues. *Nature*. **550**, 204–213 (2017).

72. D. A. Knowles *et al.*, Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods*. **14**, 699–702 (2017).

73. X. Lian *et al.*, Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling. *PNAS*. **109** (2012),

doi:10.1073/pnas.1200250109.

74. A. Liberzon *et al.*, The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

75. K. Kubara *et al.*, Status of KRAS in iPSCs Impacts upon Self-Renewal and Differentiation Propensity. *Stem Cell Reports*. **11**, 380–394 (2018).

76. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*. **12** (2015), doi:10.1038/nmeth.3582.

77. J. Ernst, M. Kellis, Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).

78. R. E. Consortium *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

79. M. A. Burke, S. A. Cook, J. G. Seidman, C. E. Seidman, Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy. *J. Am. Coll. Cardiol.* **68**, 2871–2886 (2016).

80. K.-W. Hong *et al.*, Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians. *Hum. Mol. Genet.* **23**, 6659–6667 (2014).

81. D. E. Arking *et al.*, Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat. Genet.* **46**, 826–836 (2014).

82. M. B. Rook, M. M. Evers, M. A. Vos, M. F. A. Bierhuizen, Biology of cardiac sodium channel Nav1.5 expression. *Cardiovasc. Res.* **93**, 12–23 (2011).

83. Z. Zhou *et al.*, ZNF606 interacts with Sox9 to regulate chondrocyte differentiation. *Biochem. Biophys. Res. Commun.* **479**, 920–926 (2016).

84. C. Churchhouse, B. Neale, "Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank" (2017), (available at www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank).

85. W. J. Astle *et al.*, The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. **167**, 1415-1429.e19 (2016).

86. J. F. Degner *et al.*, DNase1 sensitivity QTLs are a major determinant of expression variation. *Nature*. **482**, 390–394 (2012).

87. M. Lázaro-Gredilla, S. Van Vaerenbergh, N. D. Lawrence, Overlapping Mixtures of Gaussian Processes for the data association problem. *Pattern Recognit.* **45**, 1386–1395 (2012).

88. J. Hensman, A. G. de G. Matthews, Z. Ghahramani, Scalable Variational Gaussian Process Classification. *J. Mach. Learn. Res.* (2015).

89. E. R. Gamazon, R. Stephanie Huang, M. Eileen Dolan, N. J. Cox, H. K. Im, Integrative genomics: Quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. *Front. Genet.* (2013), doi:10.3389/fgene.2012.00202.

90. F. Pedregosa *et al.*, Learning scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

91. G. A. Moyerbrailean *et al.*, High-throughput allele-specific expression across 250 environmental conditions. *Genome Res.* (2016), doi:10.1101/gr.209759.116.

92. B. J. Strober *et al.*, Dynamic genetic regulation of gene expression during cellular differentiation. *Science (80-.).* **364**, 1287–1290 (2019).

93. A. S. E. Cuomo *et al.*, Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* (2020), doi:10.1038/s41467-020-14457-z.

94. X. Han *et al.*, Mapping human pluripotent stem cell differentiation pathways using high throughput single-cell RNA-sequencing. *Genome Biol.* (2018), doi:10.1186/s13059-018-1426-0.

95. I. Korsunsky *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* (2019), doi:10.1038/s41592-019-0619-0.

96. L. Vallier *et al.*, Signaling pathways controlling pluripotency and early cell fate decisions of human induced pluripotent stem cells. *Stem Cells* (2009), doi:10.1002/stem.199.

97. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2019), doi:10.1038/nbt.4314.

98. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* (2008), doi:10.1088/1742-5468/2008/10/P10008.

99. J. Cao *et al.*, A human cell atlas of fetal gene expression. *Science* (2020), doi:10.1126/science.aba7721.

100. X. Han *et al.*, Construction of a human cell landscape at single-cell level. *Nature* (2020), doi:10.1038/s41586-020-2157-4.

101. A. M. Goldstein, R. M. W. Hofstra, A. J. Burns, Building a brain in the gut: Development of the enteric nervous system. *Clin. Genet.* (2013), doi:10.1111/cge.12054.

102. A. Woodhoo, L. Sommer, Development of the schwann cell lineage: From the neural crest to the myelinated nerve. *Glia* (2008), doi:10.1002/glia.20723.

103. K. K. Dey, C. J. Hsiao, M. Stephens, Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genet.* **13**, e1006599 (2017).

104. P. Carbonetto, K. Luo, K. K. Dey, J. Hsiao, M. Stephens, fastTopics: fast algorithms for fitting topic models and non-negative matrix factorizations to count data. *R Packag.*

version 0.4-11., (available at https://github.com/stephenslab/fastTopics).

105. J. Jerber *et al.*, Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *bioRxiv* (2020).

106. G. E. Hoffman, E. E. Schadt, variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* (2016), doi:10.1186/s12859-016-1323-z.

107. F. A. Wolf *et al.*, PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* (2019), doi:10.1186/s13059-019-1663-x.

108. S. J. Arnold, U. K. Hofmann, E. K. Bikoff, E. J. Robertson, Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and endoderm specification in the mouse. *Development* (2008), doi:10.1242/dev.014357.

109. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* (2016), doi:10.1038/nmeth.3971.

110. D. Evseenko *et al.*, Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* (2010), doi:10.1073/pnas.1002077107.

111. L. A. Buttitta, B. A. Edgar, Mechanisms controlling cell cycle exit upon terminal differentiation. *Curr. Opin. Cell Biol.* (2007), , doi:10.1016/j.ceb.2007.10.004.

112. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. **29**, 15–21 (2013).

113. H. M. Kang *et al.*, Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4042.

114. A. Auton *et al.*, A global reference for human genetic variation. *Nature* (2015), , doi:10.1038/nature15393.

115. J. W. Belmont *et al.*, The international HapMap project. *Nature* (2003), doi:10.1038/nature02168.

116. A. T. L. Lun *et al.*, EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* (2019), doi:10.1186/s13059-019-1662-y.

117. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4096.

118. C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* (2019), doi:10.1186/s13059-019-1874-1.

119. M. E. Ritchie *et al.*, Limma powers differential expression analyses for RNA-sequencing

and microarray studies. *Nucleic Acids Res.* (2015), doi:10.1093/nar/gkv007.

120. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* (2014), doi:10.1186/gb-2014-15-2-r29.

121. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* (2016), doi:10.1093/bioinformatics/btw313.

122. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* (2018), doi:10.1186/s13059-017-1382-0.

123. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* (2014), doi:10.1371/journal.pone.0098679.

124. D. Brawand *et al.*, The evolution of gene expression levels in mammalian organs. *Nature* (2011), doi:10.1038/nature10532.

125. I. G. Romero, I. Ruvinsky, Y. Gilad, Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13**, 505–516 (2012).

126. I. G. Romero *et al.*, A panel of induced pluripotent stem cells from chimpanzees : a resource for comparative functional genomics, 1–29 (2015).

127. B. J. Pavlovic, L. E. Blake, J. Roux, C. Chavarria, Y. Gilad, A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *Sci. Rep.* **8**, 15312 (2018).

128. L. E. Blake *et al.*, A comparative study of endoderm differentiation in humans and chimpanzees. *bioRxiv* (2017), , doi:10.1101/135442.

129. S. L. Prescott *et al.*, Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest Article Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest, 68–83 (2015).

130. F. Mora-Bermúdez *et al.*, Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife* (2016), doi:10.7554/eLife.18683.

131. N. E. Banovich *et al.*, Impact of regulatory variation across human iPSCs and differentiated cells. *Genome Res.* **28**, 1–17 (2018).

132. I. G. Romero, I. Ruvinsky, Y. Gilad, Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13**, 505–516 (2012).

133. C. K. Burrows, N. E. Banovich, B. J. Pavlovic, K. Patterson, Genetic Variation , Not Cell Type of Origin , Underlies the Majority of Identifiable Regulatory Differences in iPSCs, 1–18 (2016).

134. G. E. Hoffman, P. Roussos, Dream: Powerful differential expression analysis for repeated measures designs. *bioRxiv* (2018), , doi:10.1101/432567.

135. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* (2019), doi:10.1038/s41588-018-0268-8.

136. A. J. Rahnenfuhrer, topGO: Enrichment Analysis for Gene Ontology. **R package** (2020).

137. M. Ashburner *et al.*, Gene ontology: Tool for the unification of biology. *Nat. Genet.* (2000), , doi:10.1038/75556.

138. *et al.*, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* (2021), doi:10.1093/nar/gkaa1113.

139. A. Guemez-Gamboa, N. G. Coufal, J. G. Gleeson, Primary Cilia in the Developing and Mature Brain. *Neuron* (2014), , doi:10.1016/j.neuron.2014.04.024.

140. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* (2017), doi:10.1038/ncomms14049.

141. B. J. Pavlovic, L. E. Blake, J. Roux, C. Chavarria, Y. Gilad, A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. *bioRxiv*, 289942 (2018).

142. R. Blekhman, A database of orthologous exons in primates for comparative analysis of RNA-seq data. *Nat. Preced.* (2012), doi:10.1038/npre.2012.7054.1.

143. M. Wang, L. Kong, pblat: A multithread blat algorithm speeding up aligning sequences to genomes. *BMC Bioinformatics* (2019), doi:10.1186/s12859-019-2597-8.

144. A. Wagner, A. Regev, N. Yosef, Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* (2016), , doi:10.1038/nbt.3711.

145. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Res.* (2015), , doi:10.1101/gr.190595.115.

146. Y. Ben-Menahem, *Conventionalism* (Cambridge University Press, 2006).

147. M. Stoeckius *et al.*, Large-Scale sorting of C. elegans embryos reveals dynamics of small RNA expression. *Nat. Methods*. **6**, 745–751 (2009).

148. M. Stoeckius *et al.*, Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* (2018), doi:10.1186/s13059-018-1603-1.

149. V. M. Peterson *et al.*, Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* (2017), doi:10.1038/nbt.3973.

150. A. S. Findley *et al.*, Cell type, environmental, and stochastic factors modulate genetic effects on the transcriptome. *prep* (2021).

151. D. J. P. Barker, The Developmental Origins of Adult Disease. *J. Am. Coll. Nutr*. (2004), doi:10.1080/07315724.2004.10719428.

Appendix A: Supplementary Figures and Tables

Supplementary Figures for Chapter II

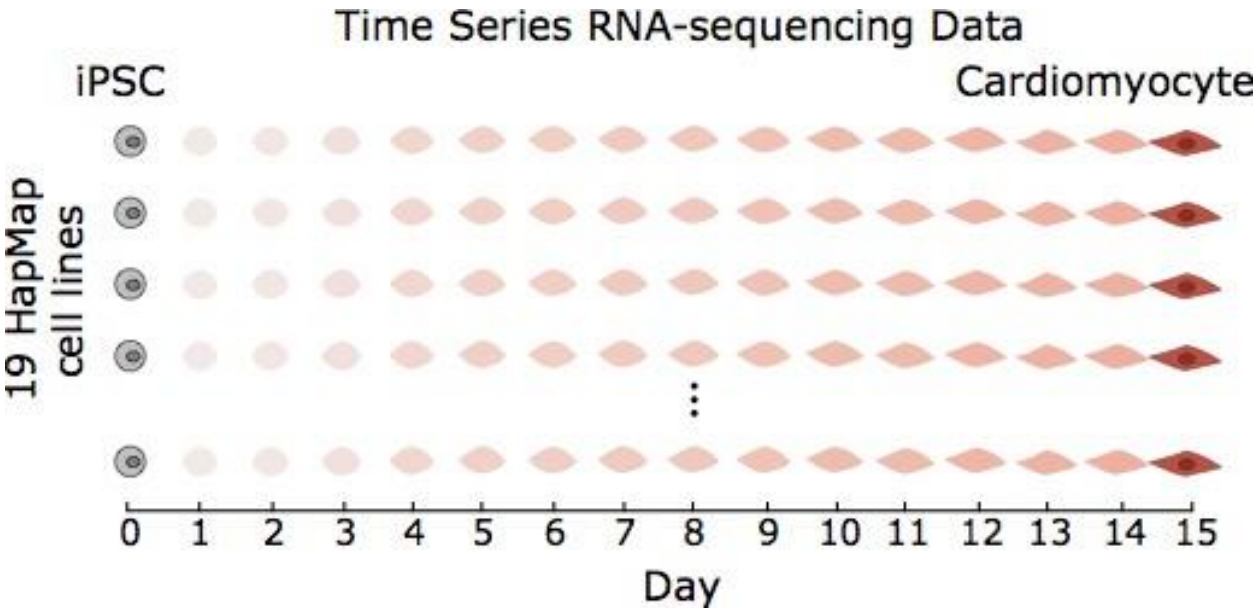


Fig. S2-1. RNA-seq sample collection. Overview of RNA-seq sample collection. In 19 Yoruba HapMap cell lines, RNA was extracted and sequenced every 24 hours at 16 time points, generating 297 RNA-seq samples.

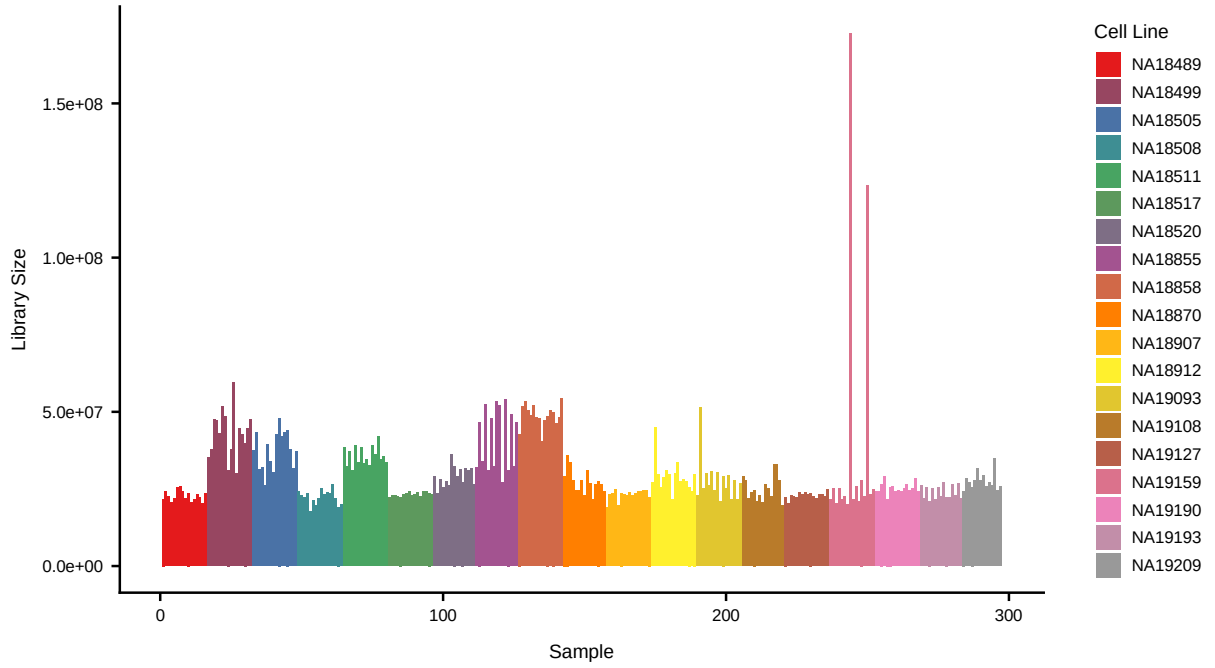


Fig. S2-2. Library size of RNA-seq samples. The library sizes of 297 RNA-seq samples colored by their cell line identity. Within each cell line, samples are ordered along the x-axis by their differentiation time point from day 0 to 15.

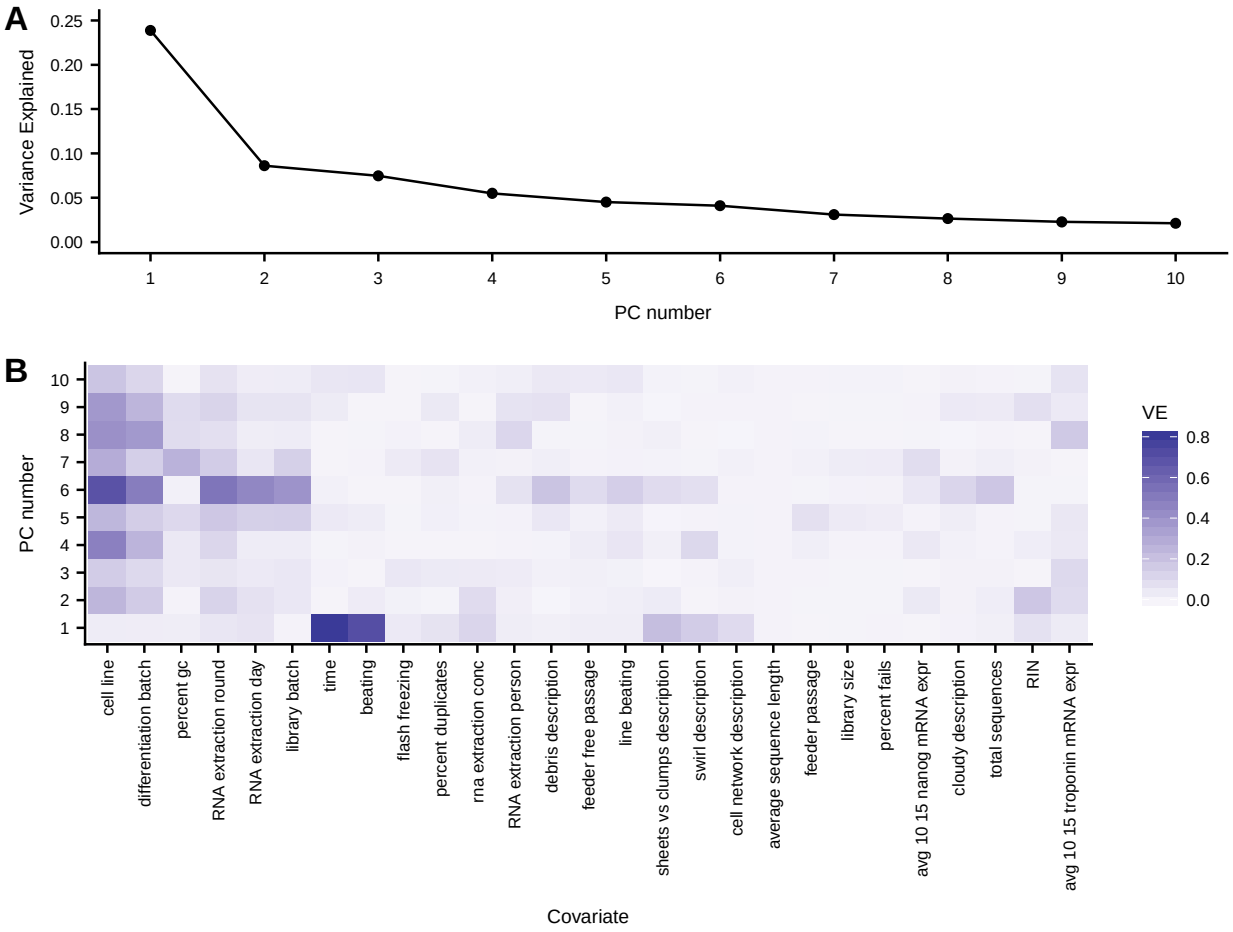


Fig. S2-3. Explaining principal components with sample covariates. (A) Variance in gene expression explained by first 10 gene expression principal components. (B) Variance explained of each gene expression principal component using sample covariates. Adjusted R^2 was used to handle categorical sample covariates. Detailed explanation of each sample covariate can be found in Table S1.

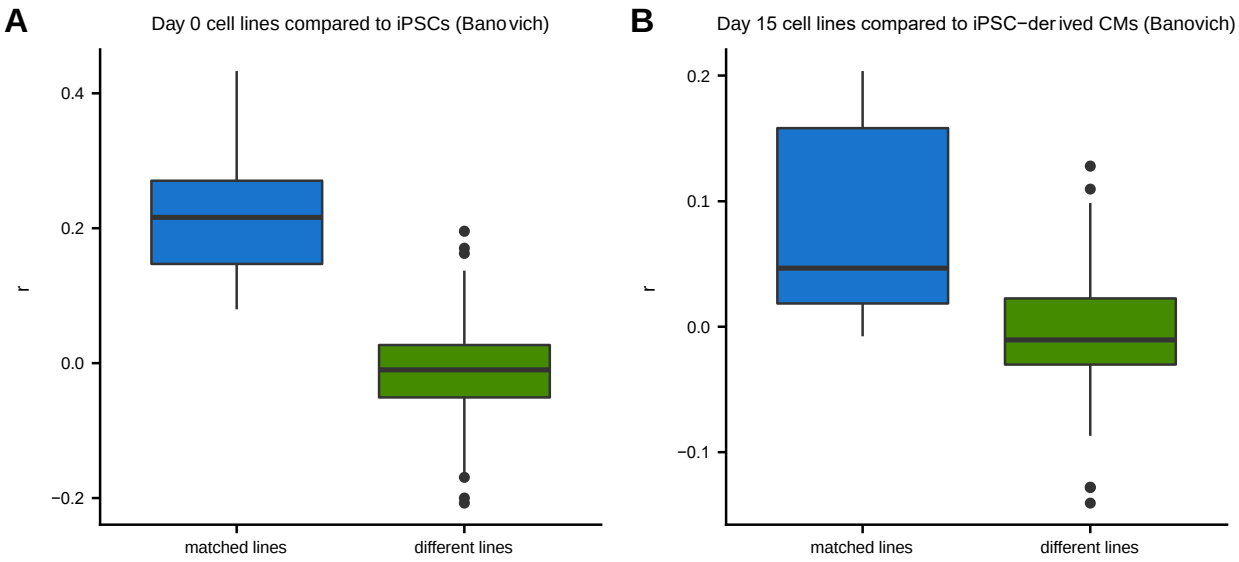


Fig. S2-4. Biological replication of day 0 and day 15 cells. We compared day 0 and day 15 cell lines with matched iPSC lines and iPSC-derived cardiomyocyte lines, respectively, from Banovich et al. (*9*). This analysis was restricted to cell lines present in both data sets. Spearman correlation across genes observed in both data sets between (A) day 0 cell lines and iPSC lines and between (B) day 15 cell lines and iPSC-derived cardiomyocyte cell lines. Distribution of spearman correlations shown for matched cell lines (blue) and different cell lines (green). The correlation of gene expression is greater for matched cell lines compared to different cell lines ($p < .05$ for both comparisons, Wilcoxon rank-sum test).

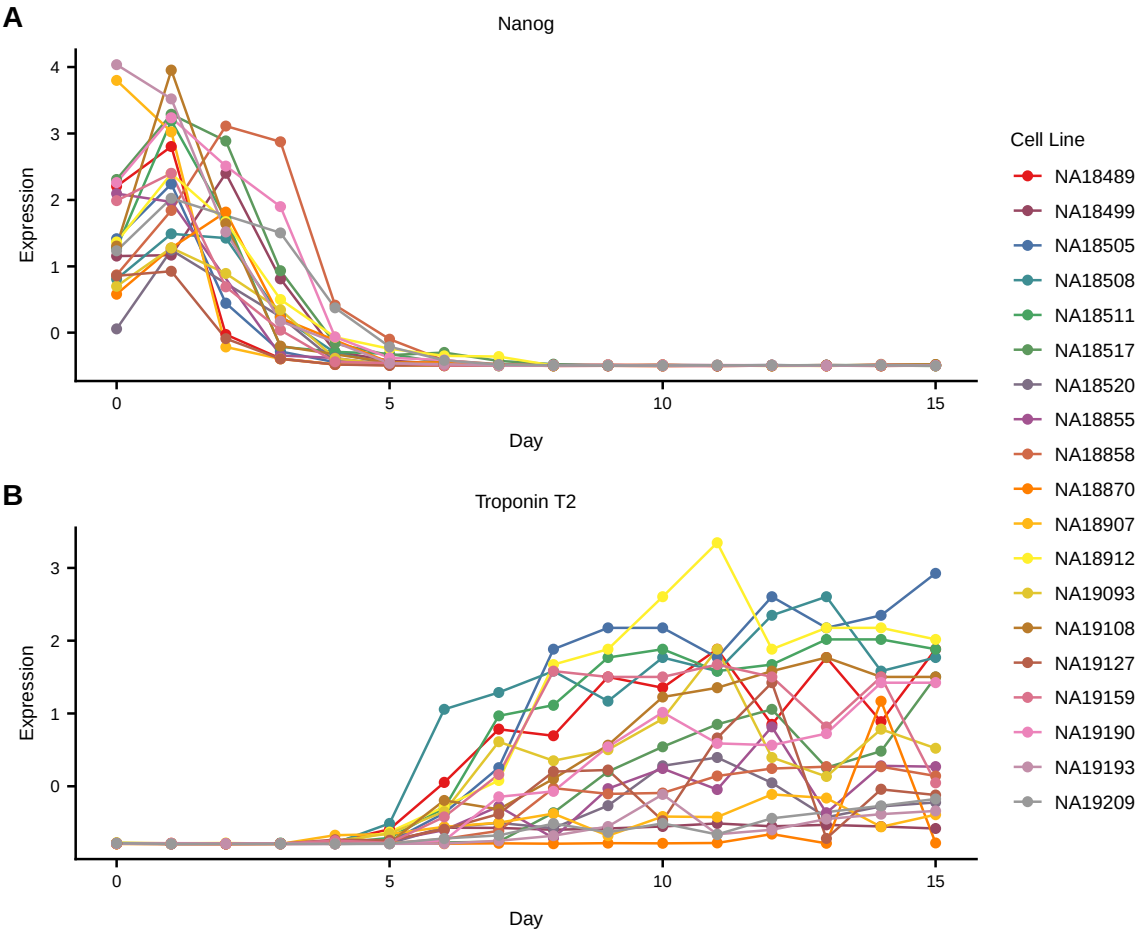


Fig. S2-5. Expression time course of known cell type specific marker genes. Standardized gene expression levels of *Nanog* (A, stem cell marker gene) and *Troponin T2* (B, cardiomyocyte marker gene) across 16 time points (x-axis) and 19 cell lines (colors).

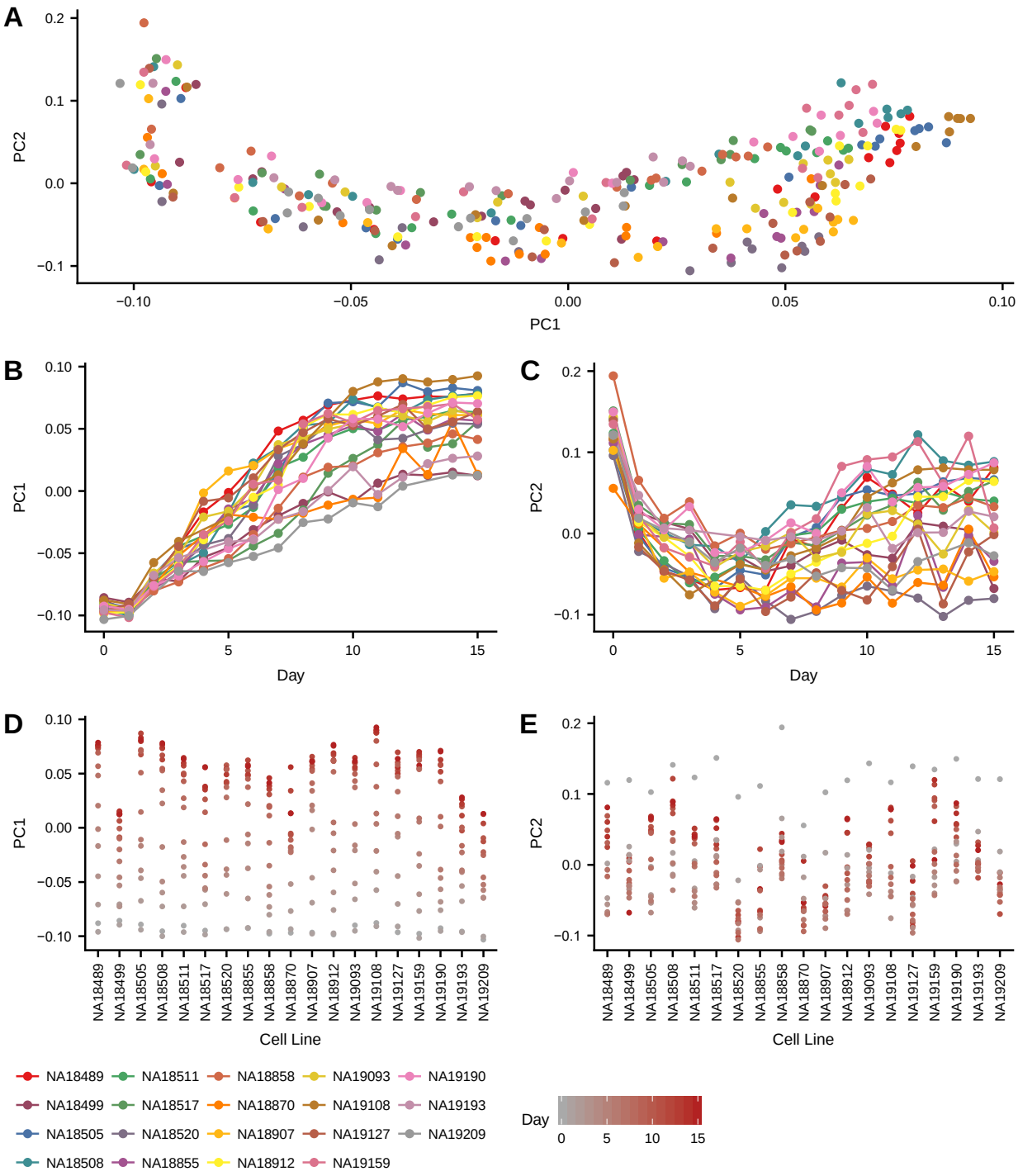


Fig. S2-6. Principal component analysis separated by cell line identity. (A) First two gene expression principal component loadings for all 297 RNA-seq samples, where each sample is colored by its cell line identity. (B, C) Principal component 1 and 2 loadings across 16 time points (x-axis) and 19 cell lines (colors). (D, E) Principal component 1 and 2 loadings across 19 cell lines (x-axis) and 16 time points (colors).

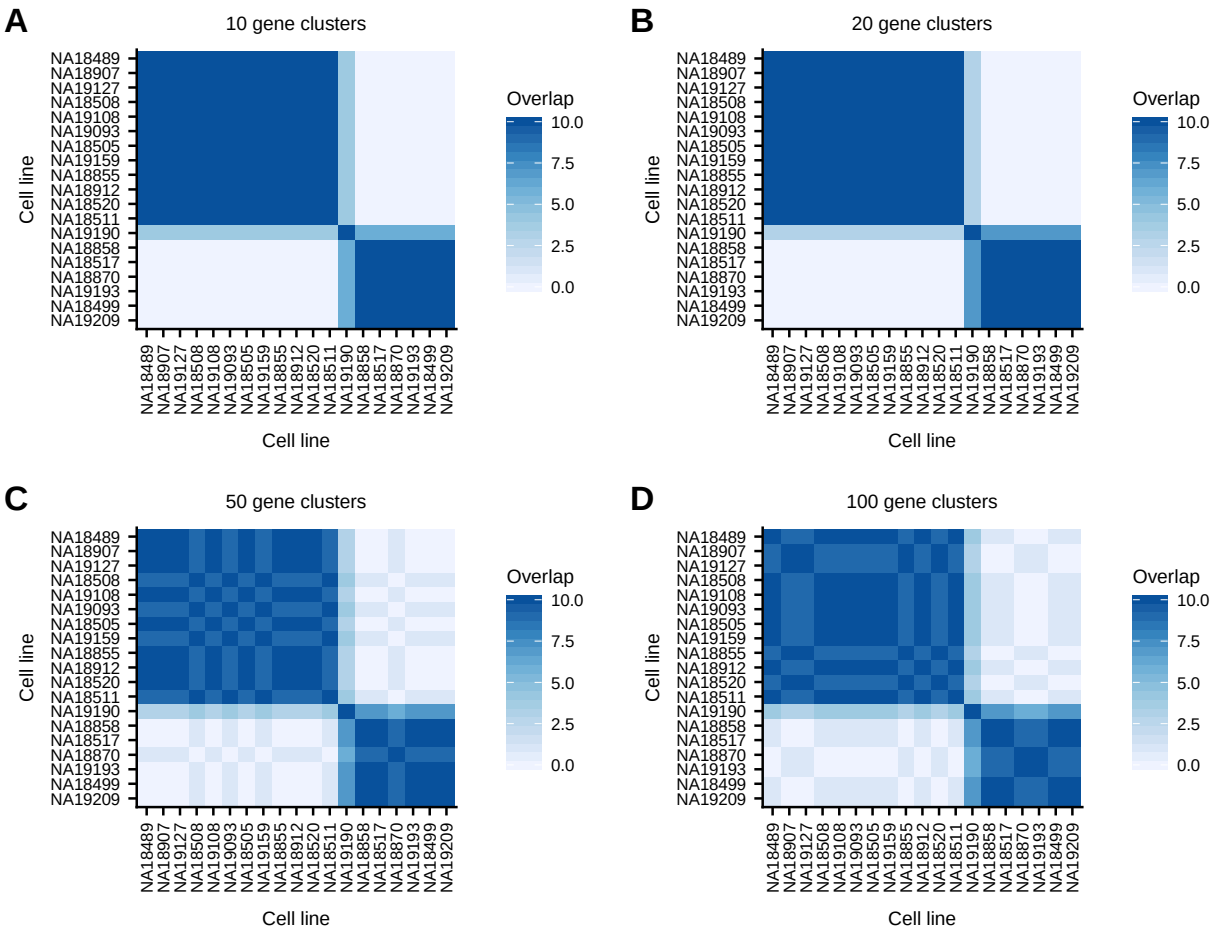


Fig. S2-7. split-GPM cell line cluster assignment robust to hyper-parameter choice. Number of times (out of 10 split-GPM runs with independent, random initializations) that each cell line pair was assigned to the same cell line cluster when 10 (A), 20 (B), 50 (C), and 100 (D) gene clusters were used. Cell lines are ordered by their cell line collapsed PC1 loadings.

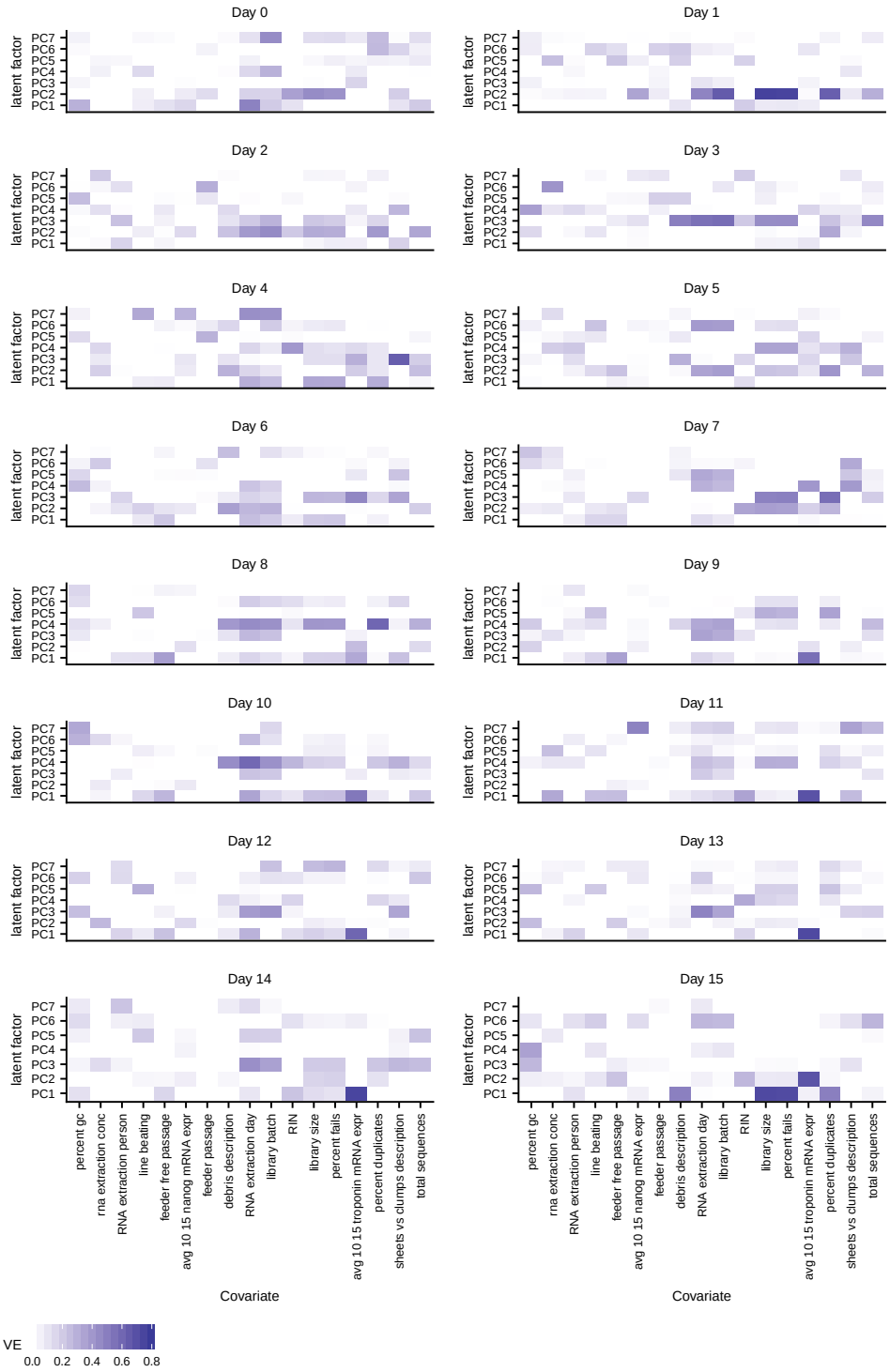


Fig. S2-8. Explaining time step principal components with sample covariates. In each time point independently, variance explained of each raw read count expression principal components (from samples belonging to the corresponding time point) using sample covariates. Adjusted R^2 was used to handle categorical sample covariates. Sample categorical covariates with more than 8 categories were excluded from this analysis due to the small sample size when considering time points, independently. Detailed explanation of each sample covariate can be found in Table S1.

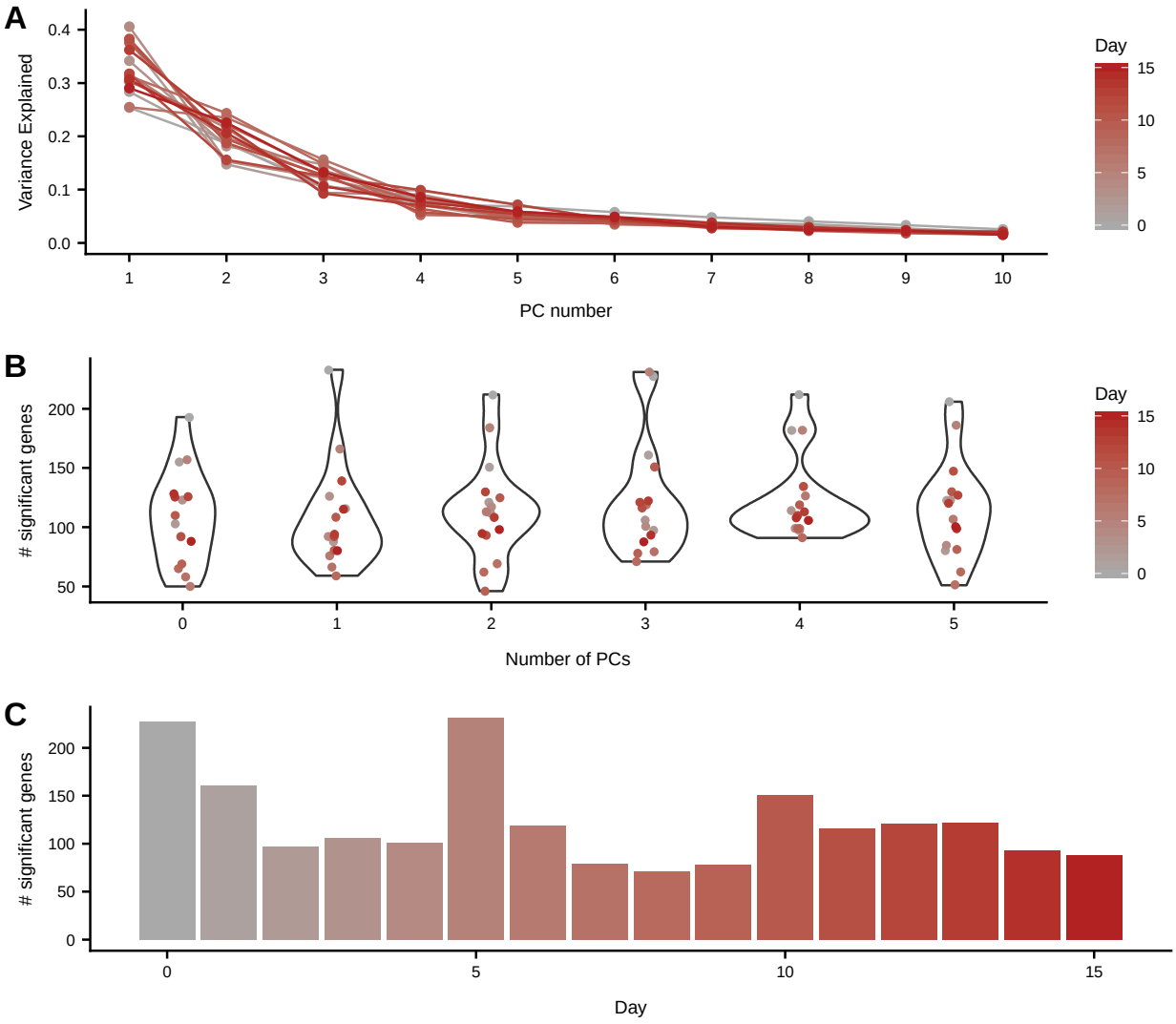


Fig. S2-9. Number of genes with non-dynamic eQTLs. (A) Variance explained of gene expression from samples belonging to a particular time point (color) by the first 10 gene expression PCs (x-axis) computed on samples belonging to that time point. (B) The number of genes with a significant eQTL (eFDR <= .05) in each time point (color) as a function of number of expression PCs controlled for (x-axis). (C) The number of genes with a significant eQTL (eFDR <= .05) in each time point when controlling for three expression PCs.

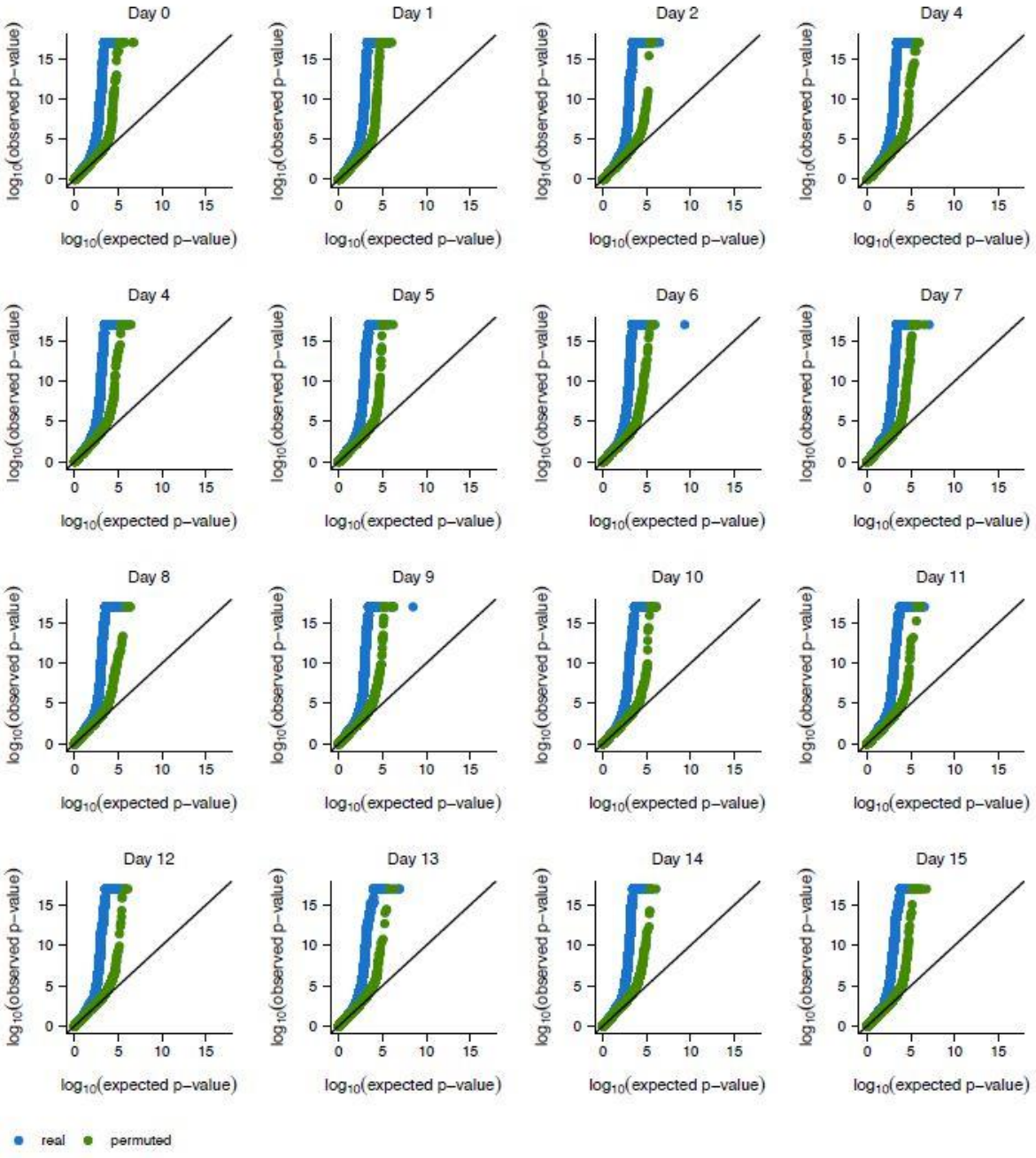


Fig. S2-10. Q-Q plots for non-dynamic eQTLs. Q-Q plot for non-dynamic eQTLs in all 16 time steps. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data (using WASP's permutation strategy) relative to uniformly distributed p-values.

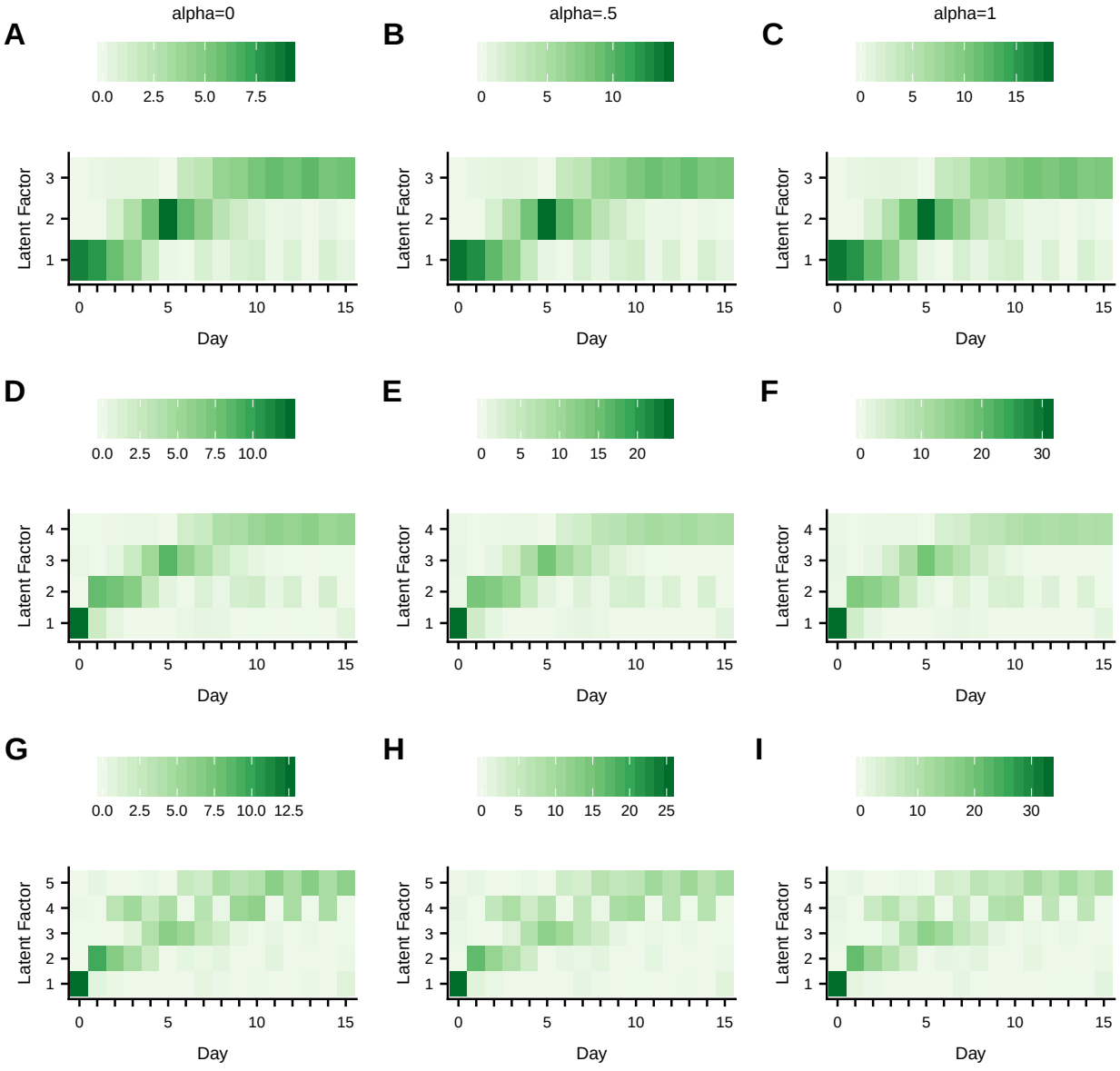


Fig. S2-11. Matrix factorization of eQTL summary statistics. Latent factors identified via sparse non-negative matrix factorization of non-dynamic eQTL -$\log_{10}$ p-values shown for a range of sparse prior choices (alpha; columns) when using 3, 4, and 5 factors (rows).

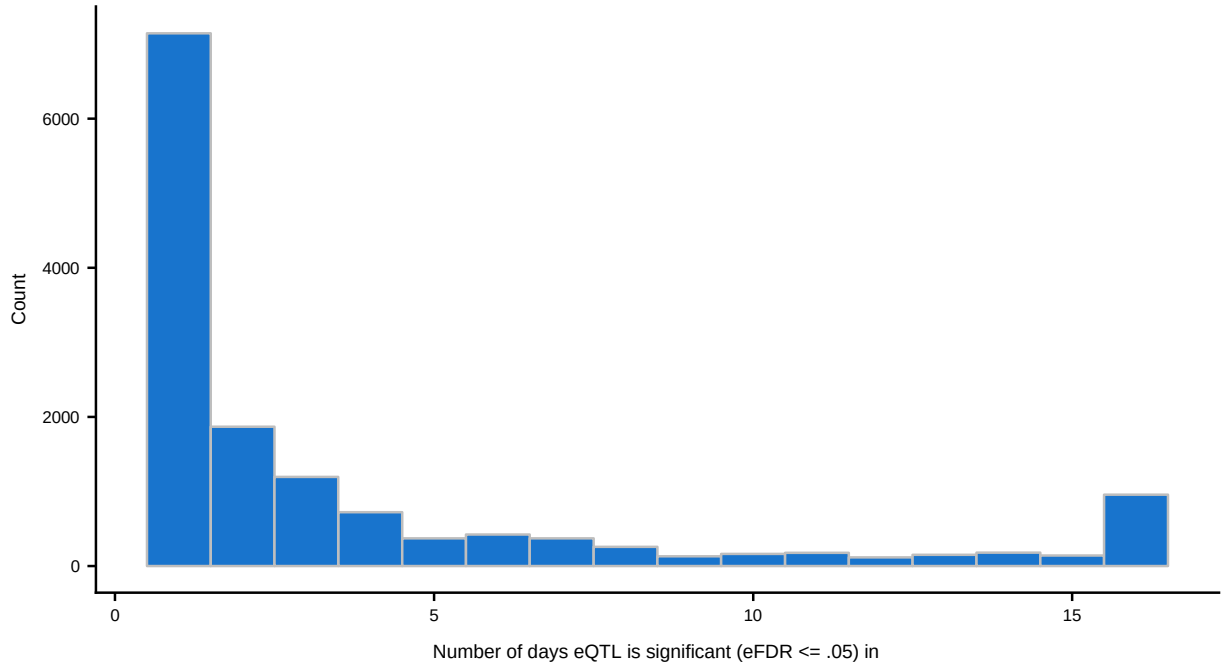


Fig. S2-12. eQTL sharing across time points. The number of days in which each non-dynamic eQTL is significant (eFDR <= .05) for all variant-gene pairs that are significant in at least one day.

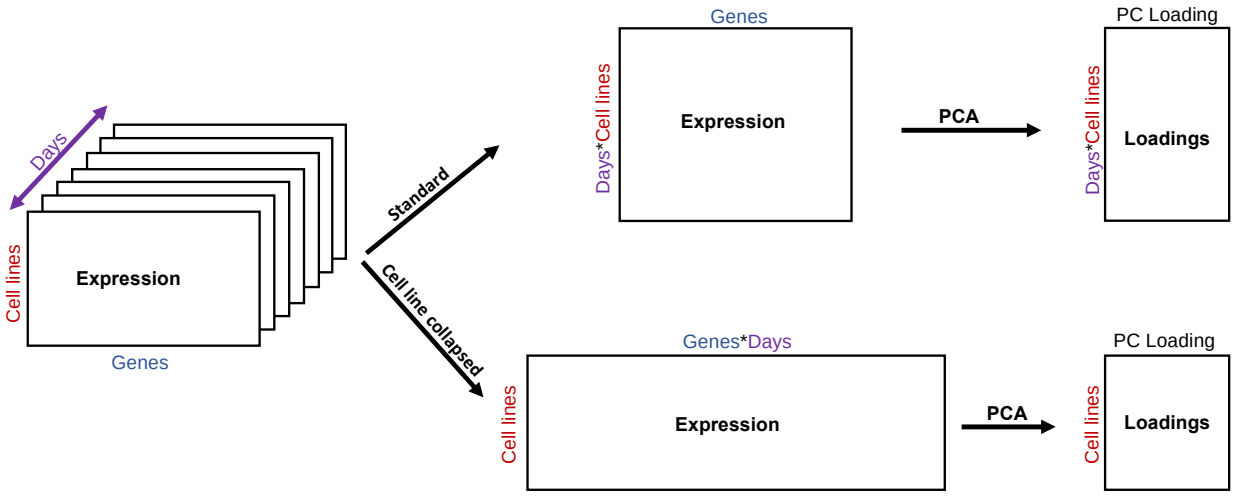


Fig. S2-13. Overview of cell line collapsed PCA. Gene expression can be represented as a three-dimensional matrix spanning days, cell lines, and genes. For standard PCA (top row), we rearrange this gene expression matrix such that rows now correspond to cell lines at specific days (e.g., RNA-seq samples) and columns correspond to genes. Here, PCA will learn a low dimensional representation for cell lines at specific days. For cell line collapsed PCA (bottom row), we rearrange this gene expression matrix such that rows now correspond to cell lines and columns correspond to genes at specific days. Here, PCA will learn a low dimensional representation for each cell line.

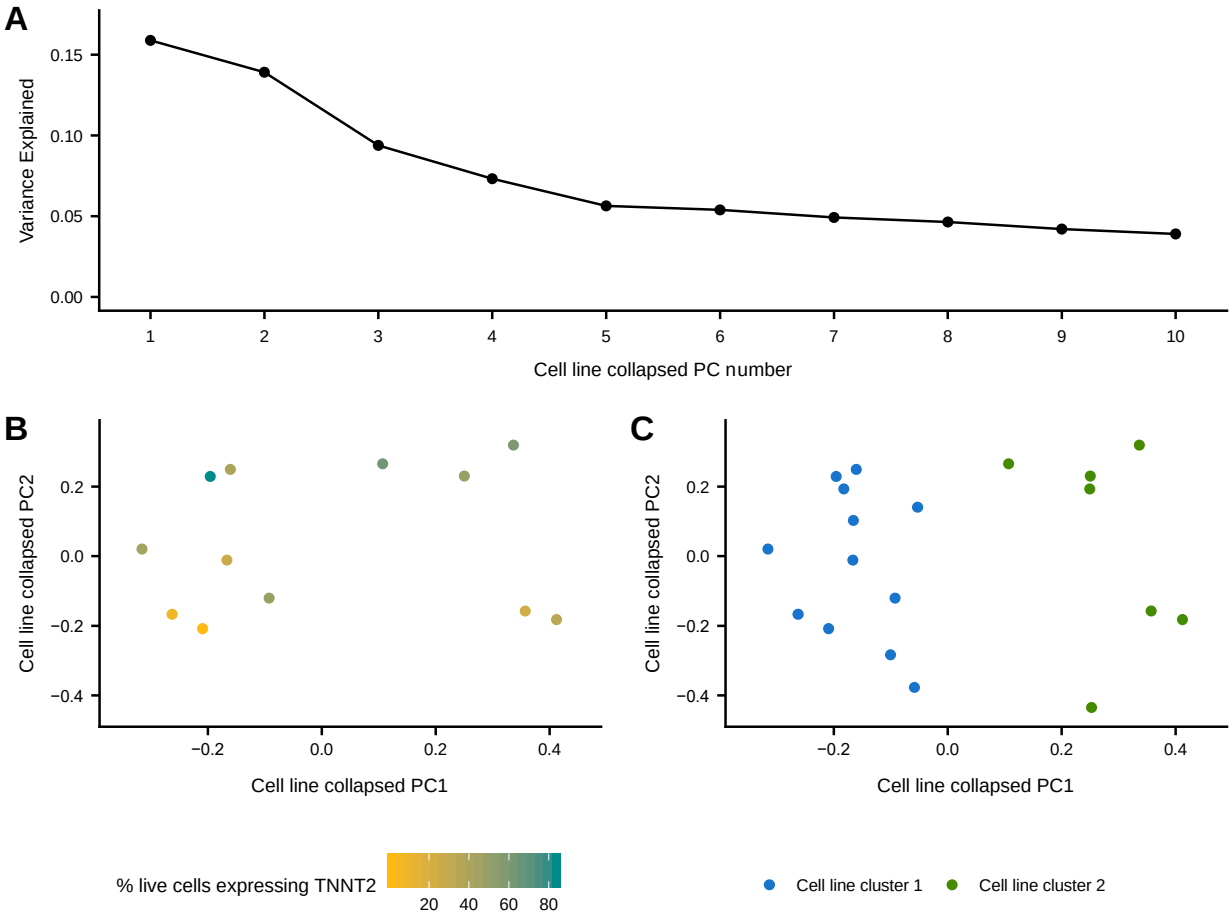


Fig. S2-14. Analysis of cell line collapsed PCs. (A) Variance explained of gene expression by first 10 cell line collapsed principal components. (B, C) First two cell line collapsed principal components where each data point is a cell line colored by its (B) percentage of live cells expressing TNNT2 at time point 15 and (C) split-GPM cell line cluster assignment.

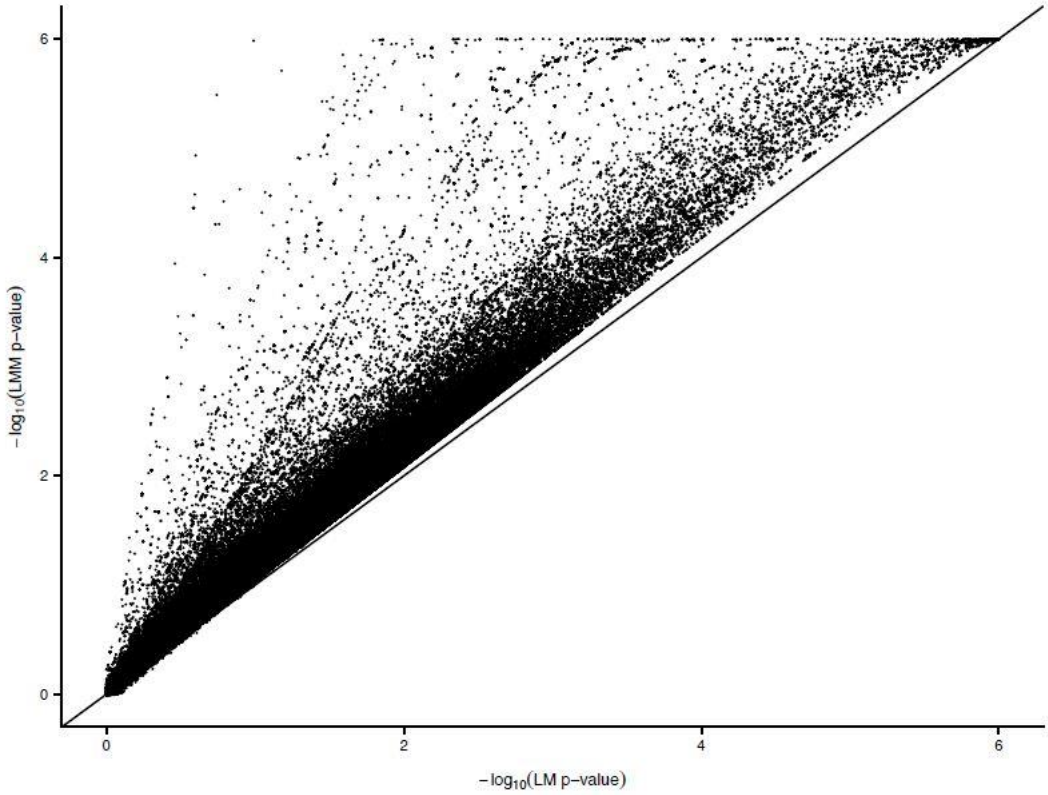


Fig. S2-15. Detecting dynamic eQTLs with gaussian linear mixed model. Comparison of linear dynamic eQTL p-values between gaussian linear model (x-axis) and gaussian linear mixed model with cell line specific random effect (y-axis) across all tested variant-gene pairs (Pearson correlation=.983).

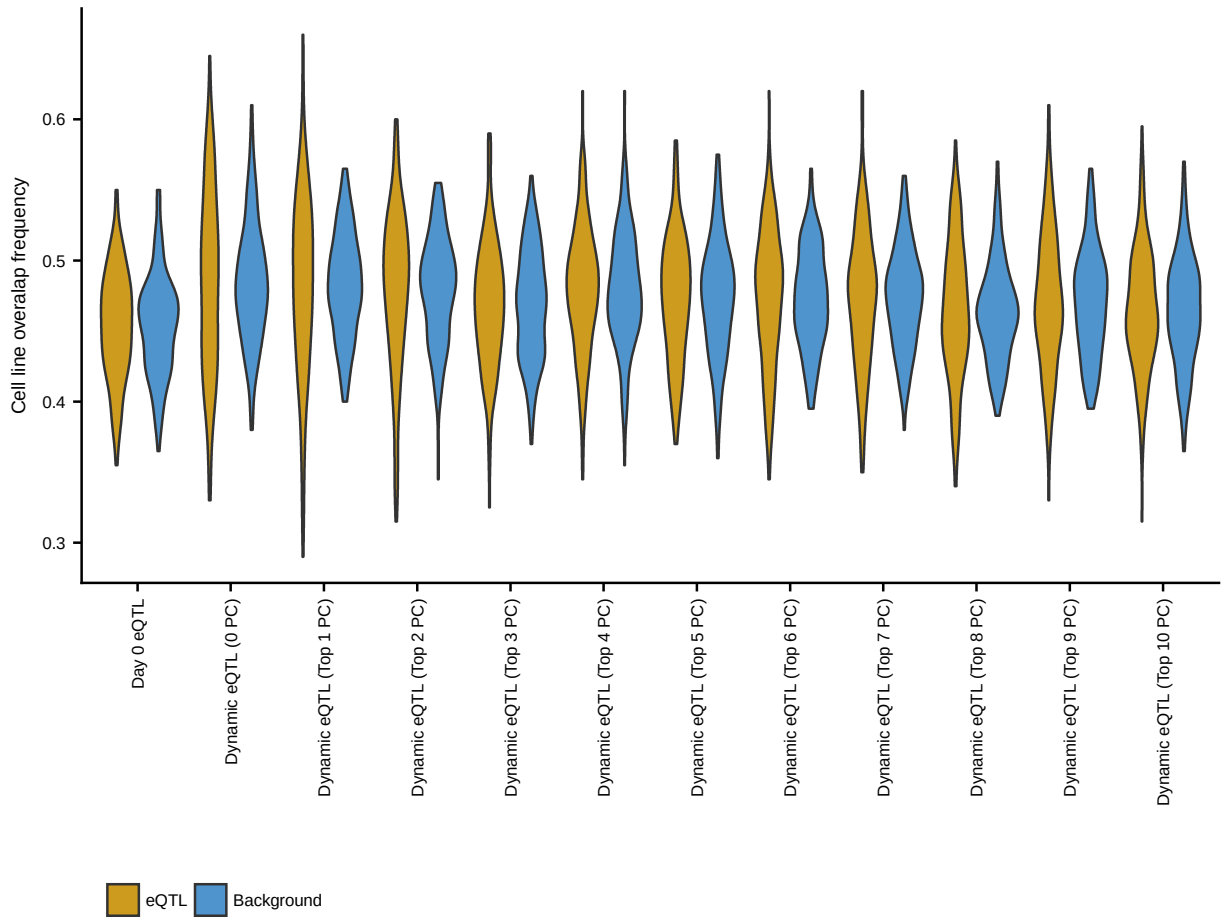


Fig. S2-16. Frequency of cell line overlap in genotype bins. Frequency at which each cell line pair is in the same genotype bin ({0,1,2}) across the strongest associated variants of the 200 most significant eQTL genes (gold) compared to MAF-matched randomly selected background variants (blue). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs. Non-dynamic eQTLs (from day 0) are also shown as a control.

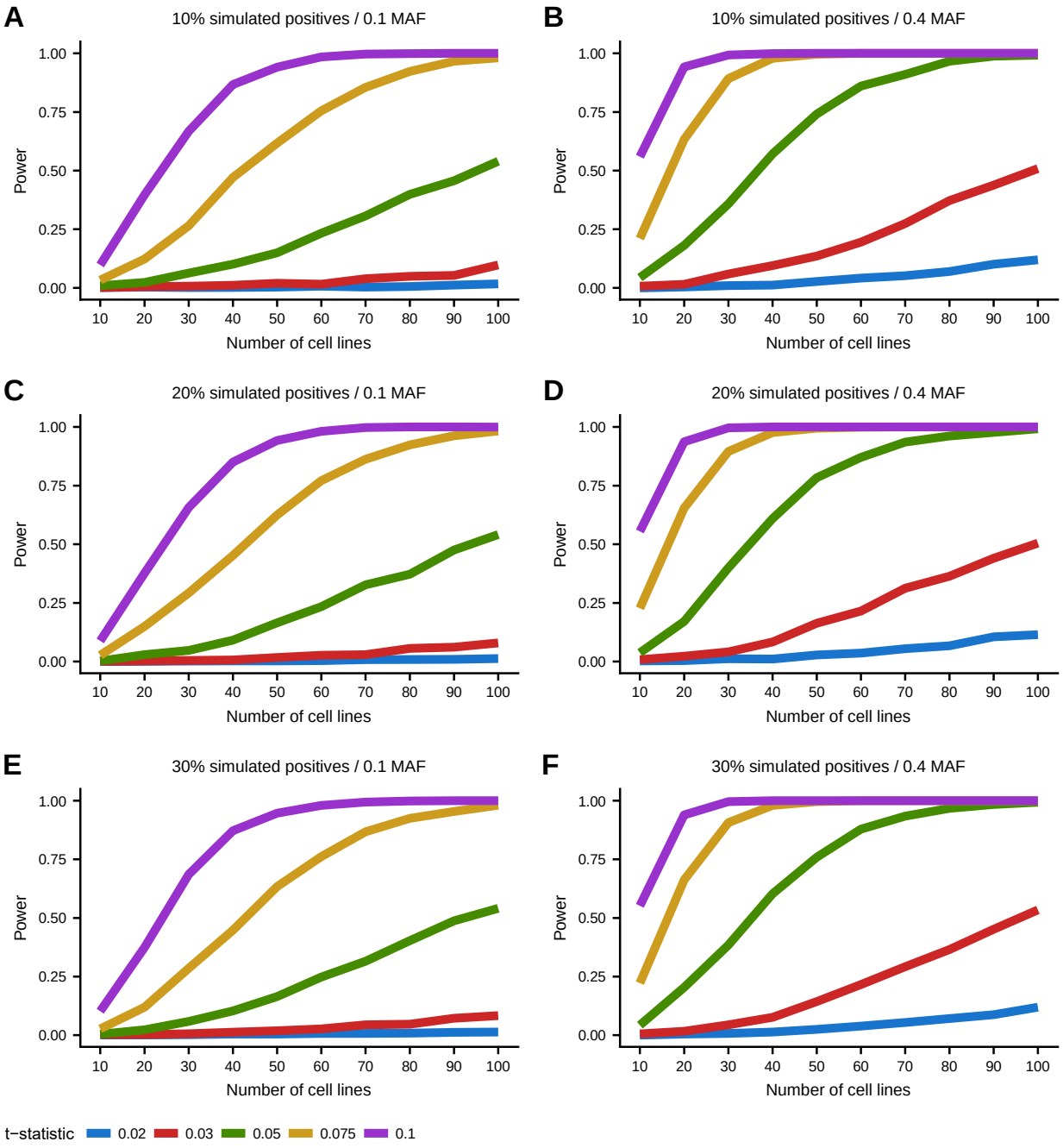


Fig. S2-17. Simulated power analysis for linear dynamic eQTLs. Power to detect simulated linear dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. We additionally vary (A-F) both the simulated MAF (columns) and the proportion of those tests that were simulated according to the alternative hypothesis (true dynamic eQTLs; rows).

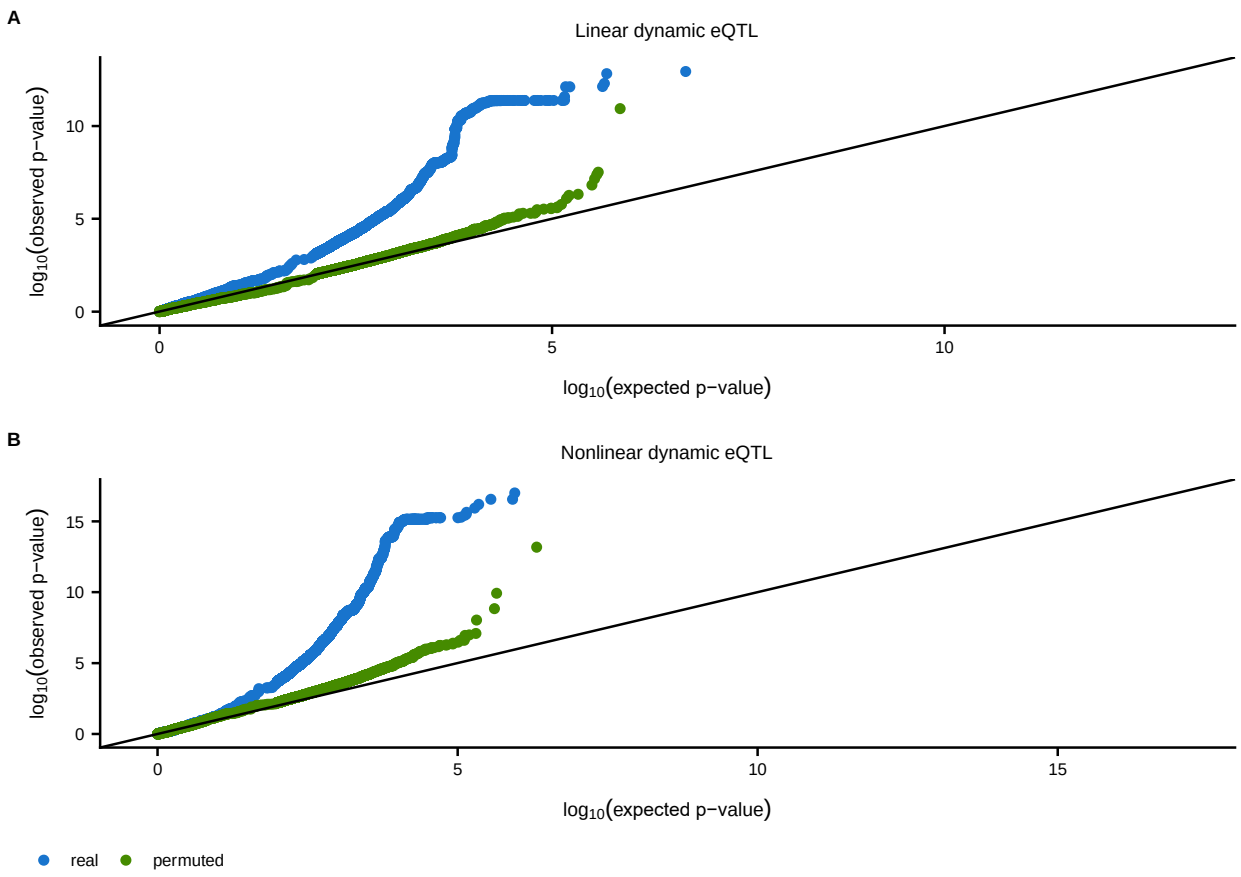


Fig. S2-18. Q-Q plots for linear and non-linear dynamic eQTLs. Q-Q plot for (A) linear and (B) non-linear dynamic eQTLs. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data relative to uniformly distributed p-values.

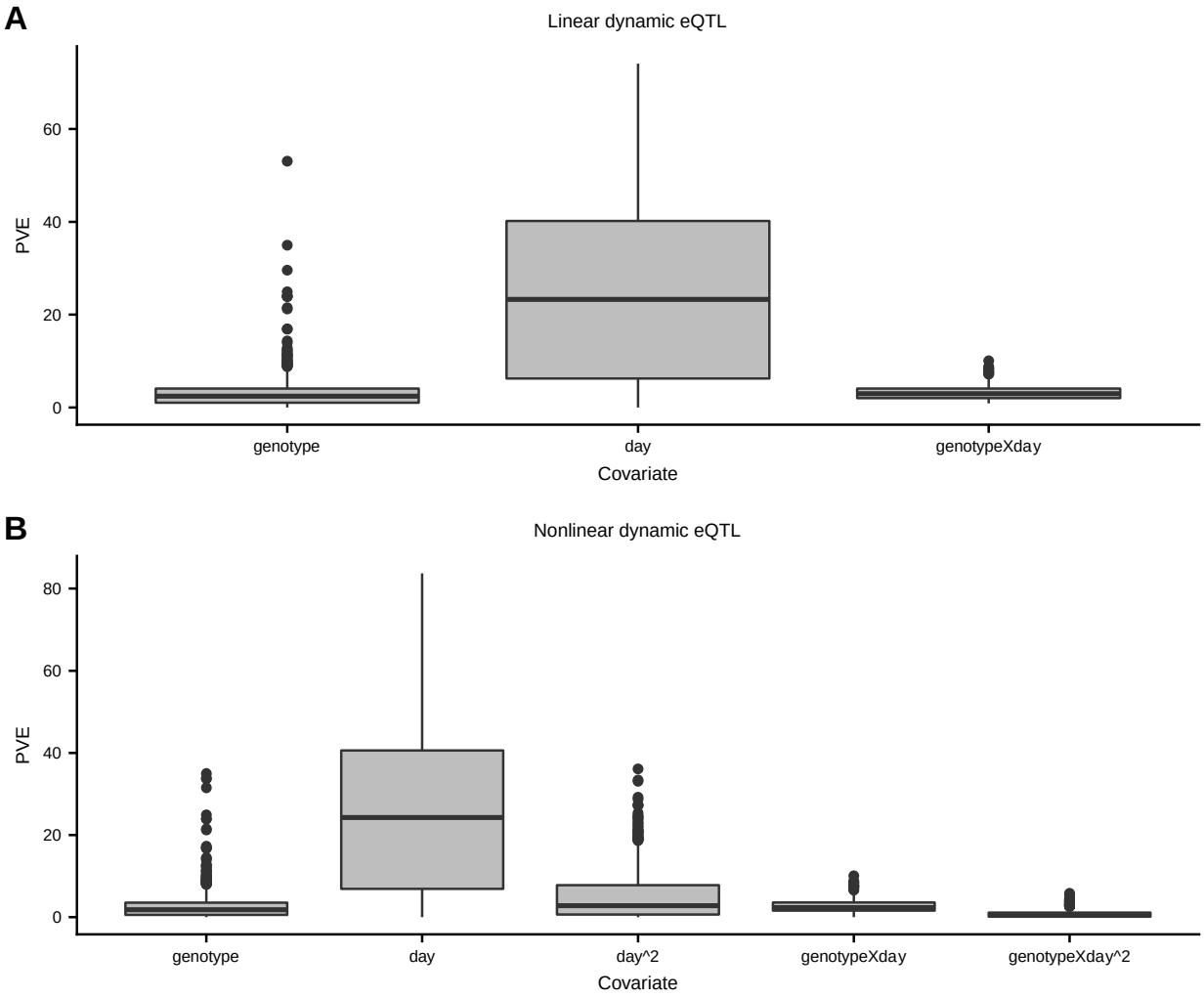


Fig. S2-19. Percent variance explained of dynamic eQTL covariates. Distribution of percent variance explained (PVE; y-axis) of each covariate (x-axis) across significant (eFDR <= .05) (A) linear dynamic eQTLs and (B) nonlinear dynamic eQTLs. For linear dynamic eQTLs, the interaction term (genotypeXday) explains on average 3.16 % of the variance. For nonlinear dynamic eQTLs, the linear interaction term (genotypeXday) and the nonlinear interaction term (genotypeXday^2) explain on average 2.69 and 0.78 % of the variance, respectively. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For linear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, and then genotypeXday. For nonlinear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, day^2, genotypeXday, and then genotypeXday^2.

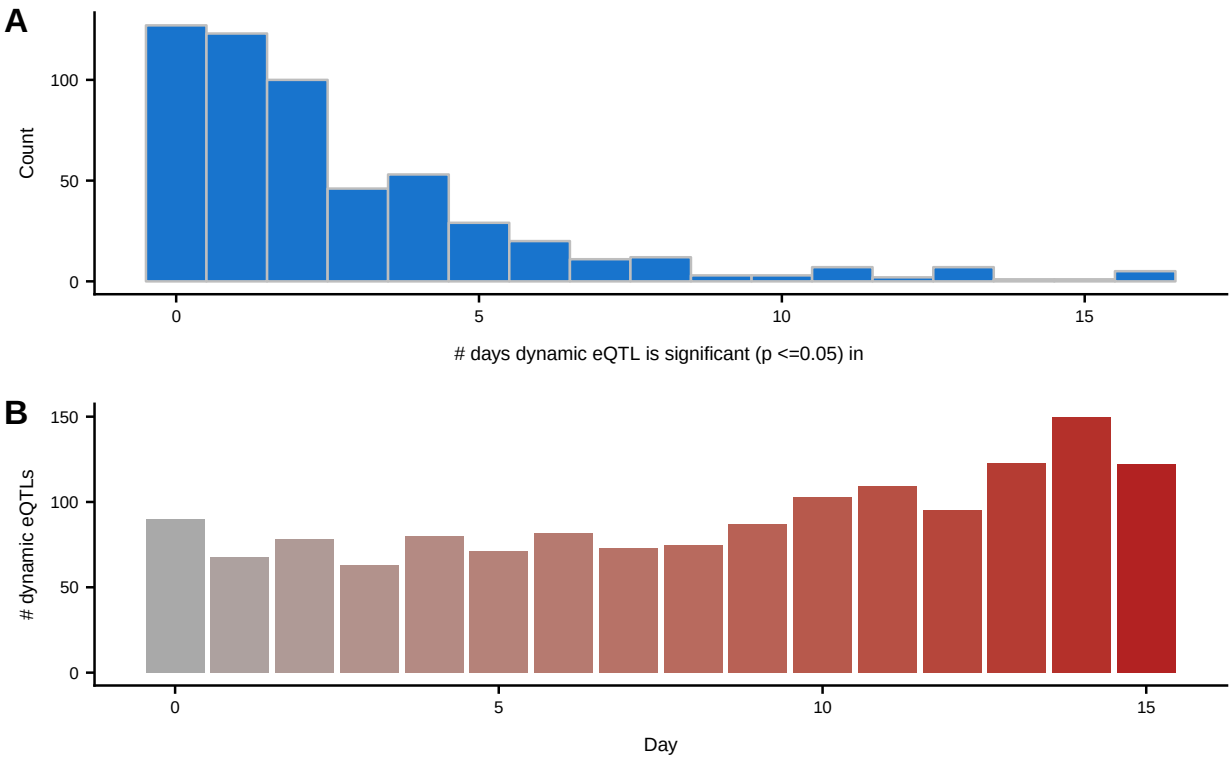


Fig. S2-20. Comparing linear dynamic eQTLs to non-dynamic eQTLs. (A) The number of time points in which the dynamic eQTLs (most significant variant per dynamic eQTL gene) have a nominally significant (p <= .05) non-dynamic eQTL. (B) The number of dynamic eQTLs (most significant variant per dynamic eQTL gene) that are nominally significant (p <= .05) in each time point.

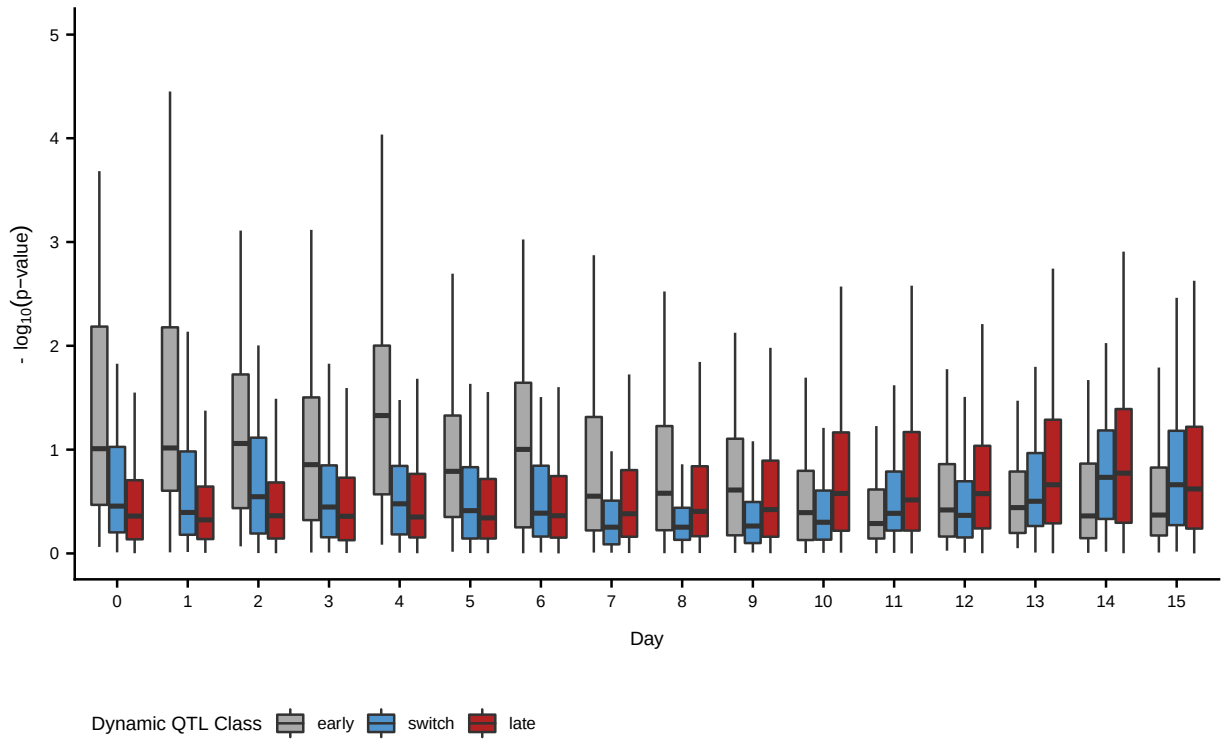


Fig. S2-21. Comparing linear dynamic eQTLs with non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of linear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by linear dynamic eQTL classifications (early, switch, and late).

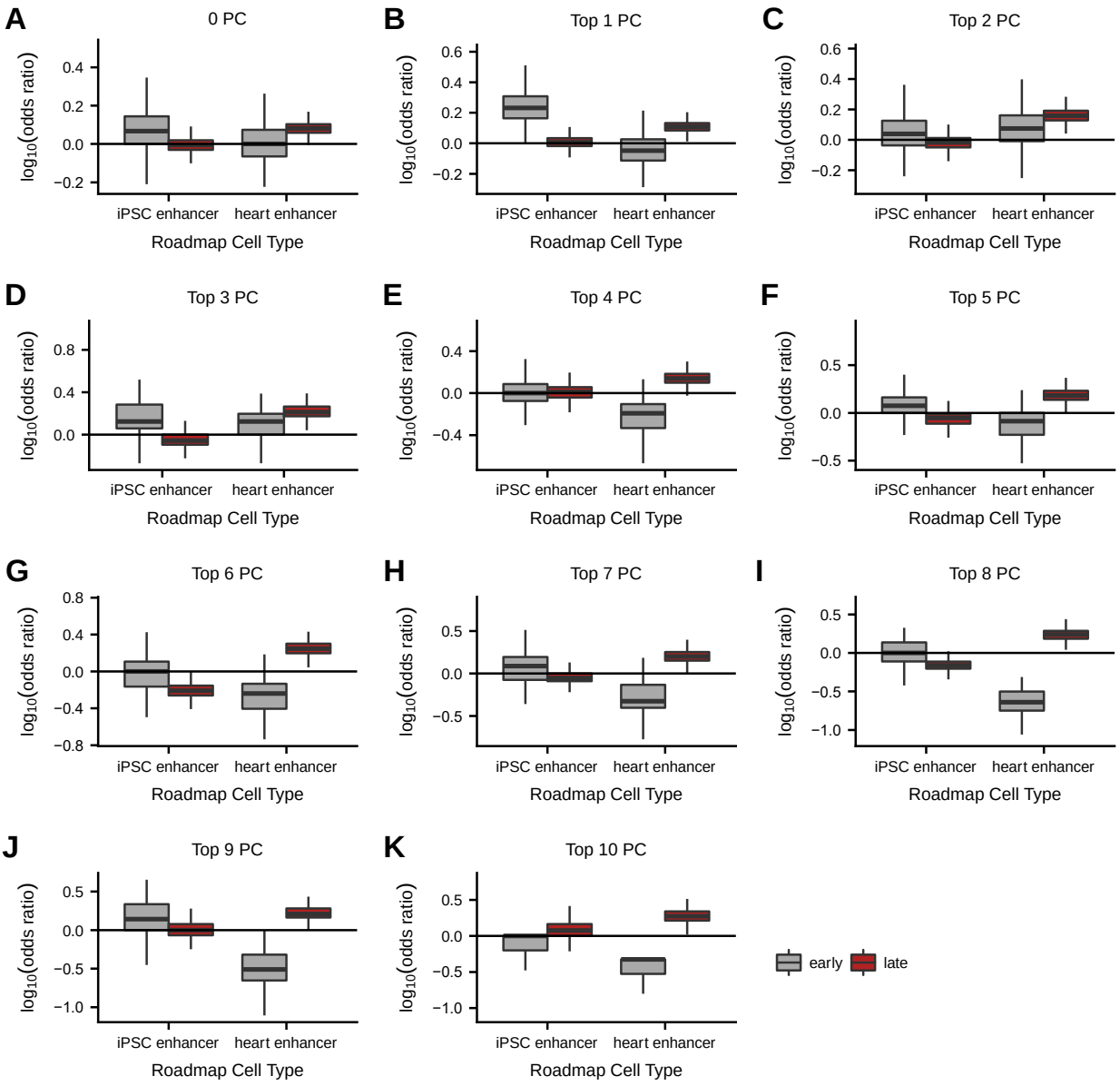


Fig. S2-22. Dynamic eQTL enhancer enrichment. Enrichment of dynamic eQTLs within cell type specific chromHMM enhancer elements relative to 1000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency. Dynamic eQTLs were classified as early (eQTL effect size decreasing over time) or late (eQTL effect size increasing over time). Analysis shown for linear dynamic eQTLs while controlling for a range of the top cell line collapsed PCs (A-K).

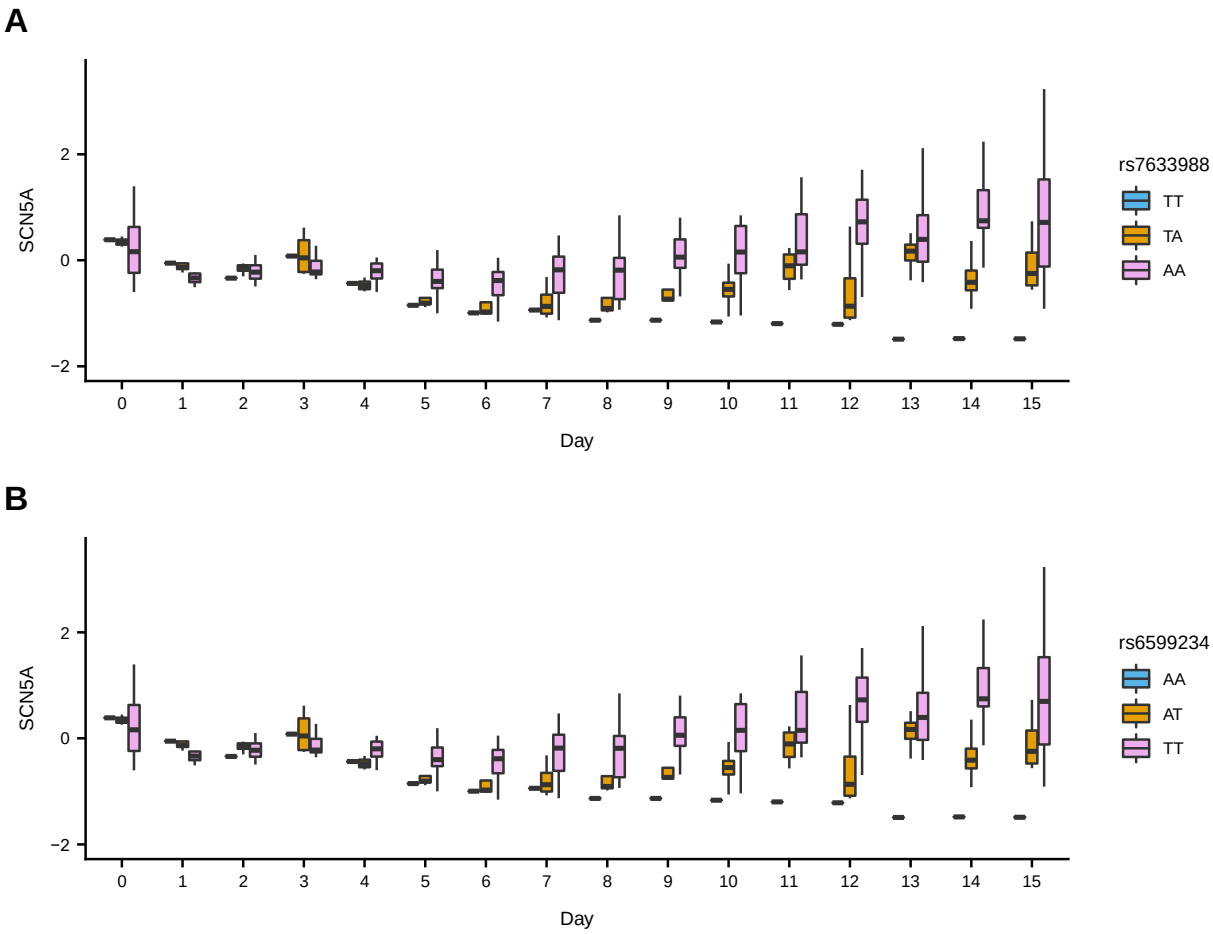


Fig. S2-23. Two significant linear dynamic eQTLs are known GWAS variants. Linear interaction association between time point (x-axis) and genotype (color) of (A) rs7633988 and (B) rs6599234 on residual gene expression (cell line effects regressed on expression) of *SCN5A* (y-axis).

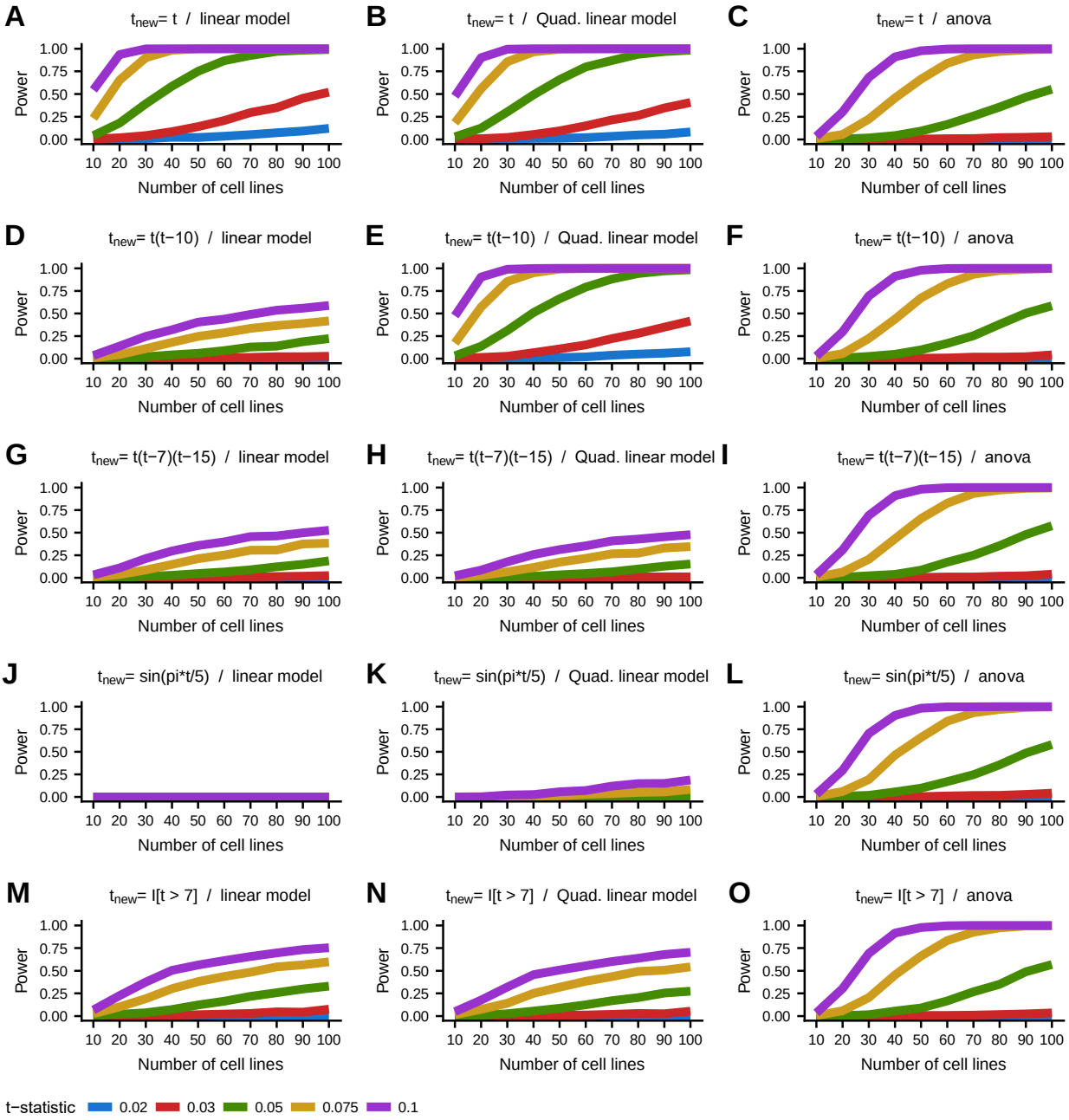


Fig. S2-24. Non-linear simulated power analysis. Power to detect simulated dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). t-statistic represents the ratio of the effect size of the interaction term and the standard deviation term used to simulate the expression data. Simulated expression was generated based on various transformations (t_{new}; rows) of the original values of differentiation time (t). Transformed differentiation time was scaled to have the same standard deviation as the original values of differentiation time. Three different statistical models were used to identify dynamic eQTLs (columns): linear model (linear dynamic eQTL), quadratic linear model (nonlinear dynamic eQTL), and categorical ANOVA analysis. The simulated MAF was .4 and 30% of all simulated tests were drawn from the alternative hypothesis.

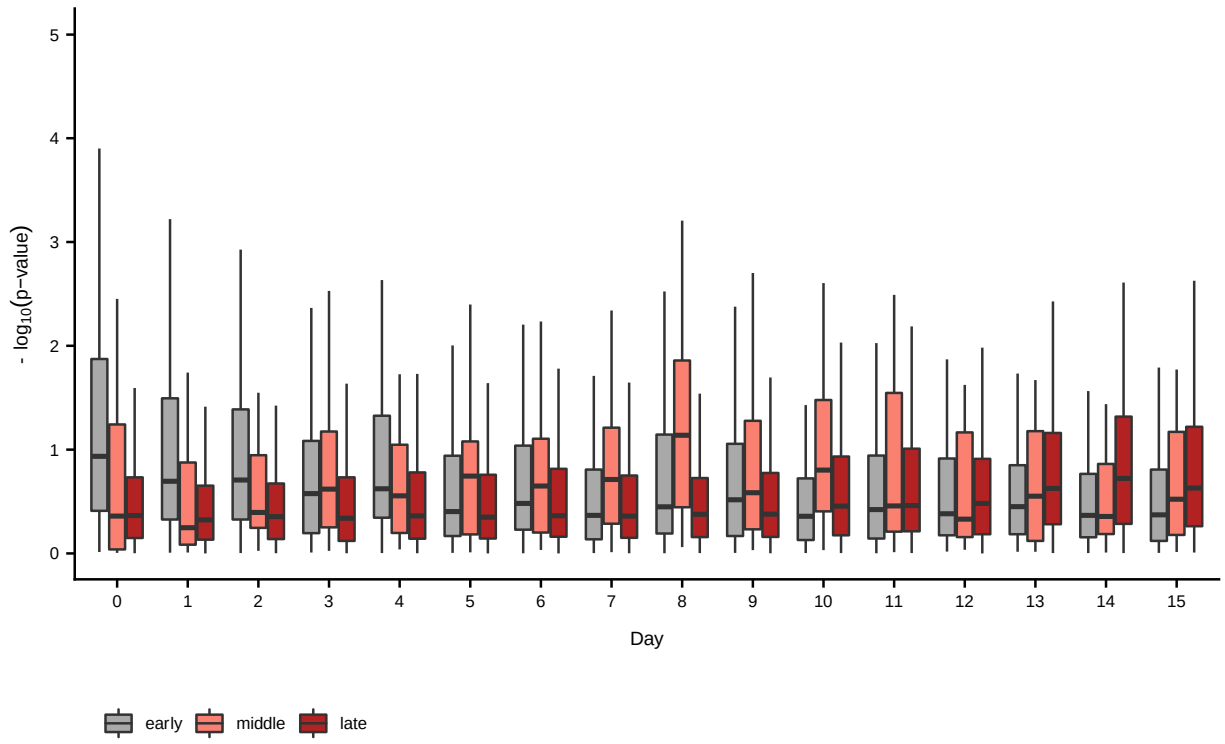


Fig. S2-25. Comparing nonlinear dynamic eQTLs to non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of nonlinear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by nonlinear dynamic eQTL classifications (early, middle, and late).

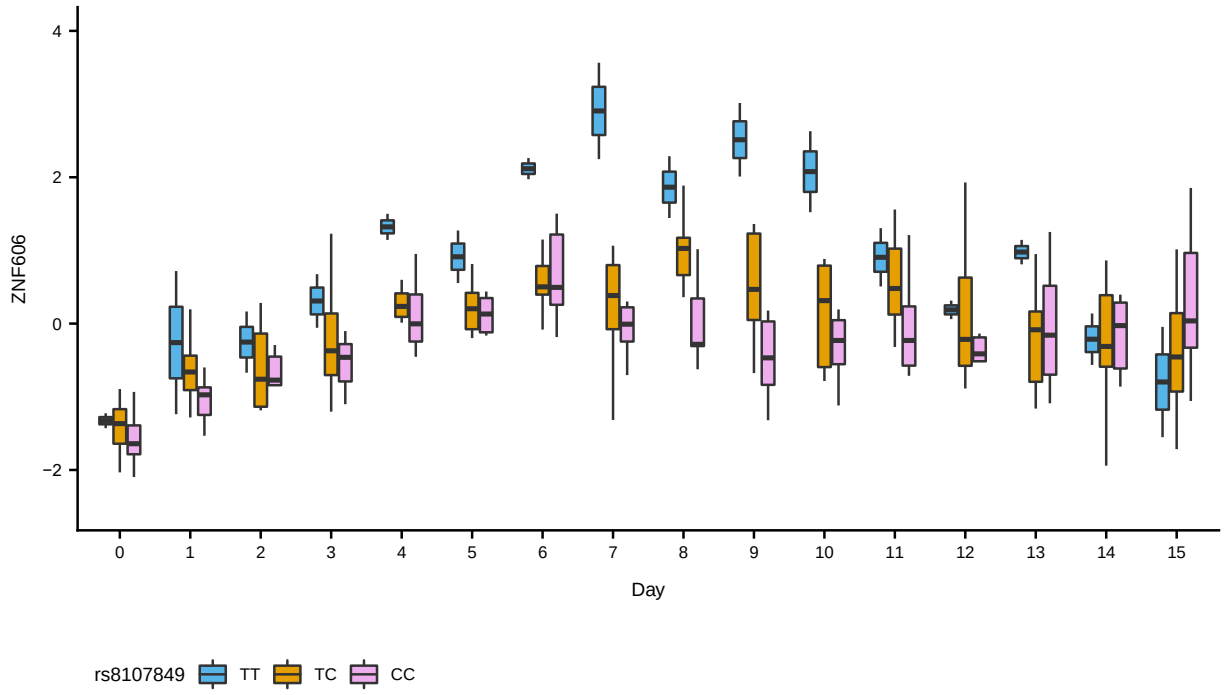


Fig. S2-26. Middle dynamic eQTL example. Nonlinear interaction association between genotype (color) of rs8107849 and time point (x-axis) on residual gene expression (cell line effects regressed on expression) of *ZNF606* (y-axis).

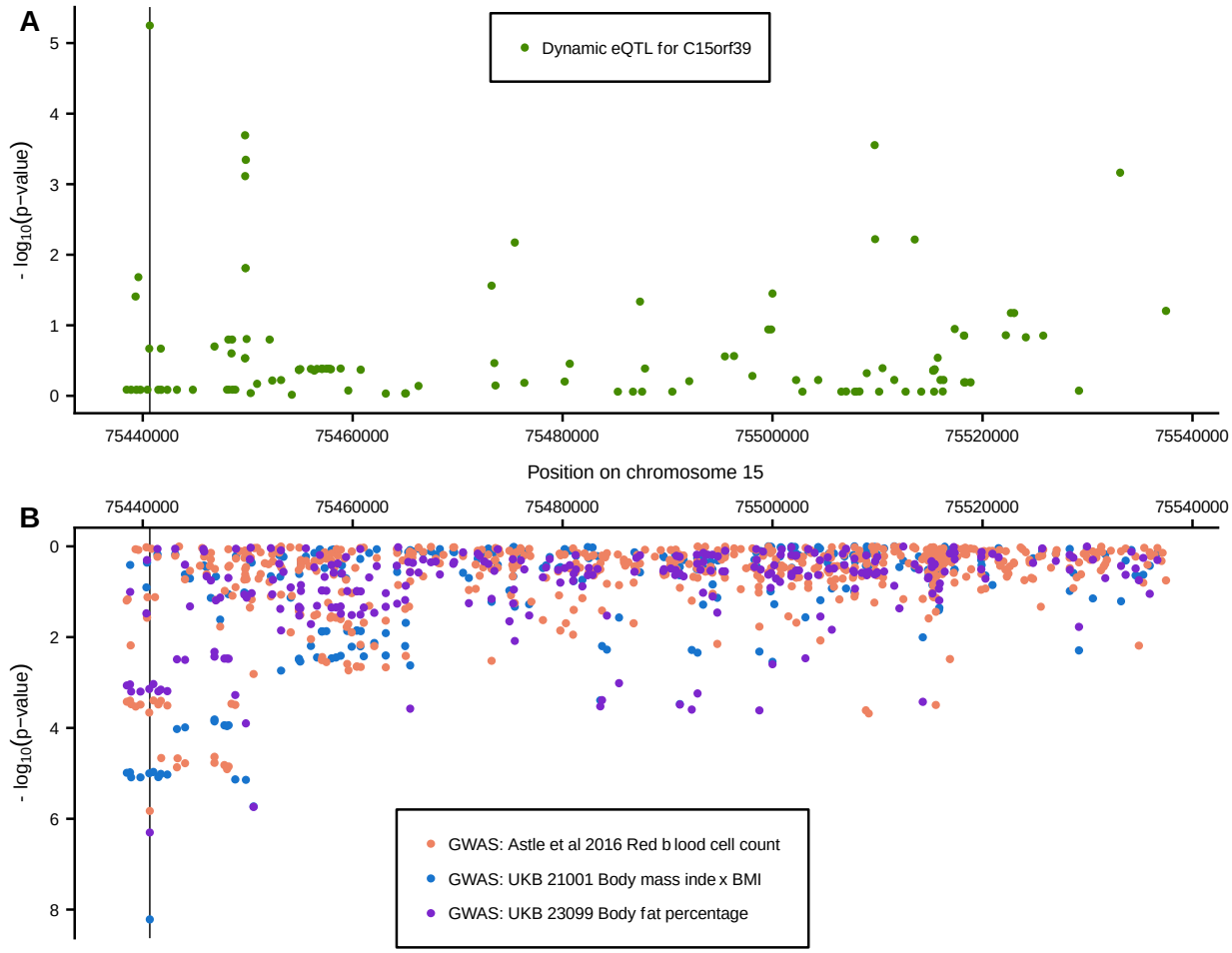


Fig. S2-27. Nonlinear dynamic eQTL overlaps GWAS variant. (A) Manhattan plot showing interaction association p-values for *C15orf39* according to nonlinear dynamic eQTL calling for all variants tested within 50KB of the *C15orf39* transcription start site. (B) Manhattan plot showing GWAS p-values on the same region surrounding *C15orf39* from three different GWAS studies (colors) (*23*, *24*). Vertical line depicts genomic location of most significant nonlinear dynamic eQTL (rs28818910) for *C15orf39*. p-values shown for body mass index and body fat percentage are based on round 1 of UK Biobank (UKB) (*23*). Body mass index and body fat percentage p-values for rs28818910 according to the round 2 of UKB (*34*) become slightly less extreme (p=1.322e-07 and p= 2.521e-06, respectively), but are still significant after multiple testing correction for all significant (eFDR <= .05) nonlinear dynamic eQTL variants (Bonferroni p= 0.000902 and Bonferroni p=.0172, respectively).

Supplementary Tables for Chapter II

Table S2-1. Sample metadata. Available as an excel file online (Strober et al. 2019). Sheet ‘A-Sample meta-data’ contains meta-data for each RNA-seq sample. Sheet ‘B-meta data description’ contains descriptions of each meta-data variable collected.

Cell Line	Percent of Live Cells Expressing TNNT2
18489	44.3
18499	24.2
18505	NA
18508	83.9
18511	NA
18517	47.8
18520	NA
18855	NA
18858	NA
18870	NA
18907	7.9
18912	47.8
19093	27
19108	NA
19127	1.1
19159	39.8
19190	63.2
19193	59.5
19209	33.4

Table S2-2. Flow cytometry results for each cell line at day 15 of cardiomyocyte differentiation. The percent of live cells expressing cardiac troponin (TNNT2) for every cell line at day 15 of differentiation. Cells with an NA indicate that flow cytometry was not performed on this cell line.

Hallmark gene set	Gene cluster 2	Gene cluster 4	Gene cluster 5	Gene cluster 6	Gene cluster 9	Gene cluster 11	Gene cluster 13	Gene cluster 16
TNFA signaling via NFKB	1	1	1	.000208	1	1	1	1
Mitotic spindle	1	1	1	1	.0166	1	1.80e-14	1
TGF beta signaling	1	1	1	.348	.000624	1	1	1
DNA repair	1	1	.000242	1	1	1	3.73e-7	1
G2M checkpoint	1	1	1	1	1	1	2.87e-63	.594
Myogenesis	9.29e-14	1	1	1.05e-5	1	1	1	1
Protein secretion	.00384	1	1	1	1	1	1	1
Complement	1	1.98e-5	1	1	1	1	1	1
Unfolded protein response	1	1	6.99e-5	1	1	1	1	1
MTORC1 signaling	1	1	2.07e-10	1	1	.696	1	1
E2F targets	1	1	.0111	1	1	1	5.47e-73	.0458
MYC targets V1	1	1	3.03e-25	1	1	.329	1.28e-16	1.16e-5
MYC targets V2	1	1	7.04e-21	1	1	1	.981	1
Epithelial mesenchymal transition	1	.000310	1	2.05e-5	1	1	1	1
Xenobiotic metabolism	1	.000435	1	1	1	1	1	1
Oxidative phosphorylation	1	1	1	.134	1	8.11e-11	1	1
Heme metabolism	1.24e-6	1	1	1	1	1	1	1
Coagulation	1	1.72e-16	1	1	1	1	1	1
Bile acid metabolism	1	.00392	1	1	1	1	1	1
Spermatogenesis	1	1	1	1	1	1	.00433	1
KRAS signaling up	1	.00536	1	.622	1	1	1	1

Table S2-3. Hallmark gene set enrichment of split-GPM gene clusters. Bonferroni corrected p-values (Fisher's exact) from gene set enrichment of gene clusters (columns) from split-GPM within Hallmark gene sets (rows). Only gene clusters and gene sets with at least one significant enrichment (Bonferroni p-value <= .05) are shown.

# of cell line collapsed PCs	# genes with significant dynamic eQTL (eFDR <= .05)	# genes with significant dynamic eQTL (eFDR <= .01)
0	2256	931
1	1943	785
2	1247	294
3	648	250
4	608	186
5	550	150
6	533	113
7	556	212
8	456	110
9	288	22
10	213	79

Table S2-4. Number of linear dynamic eQTLs detected. The number of genes with a significant linear dynamic eQTL (eFDR <= .05 and eFDR <= .01) as a function of the number cell line collapsed PCs used as covariates.

Table S2-5. Percent variance explained for linear dynamic eQTLs. Available as a text file online (Strober et al. 2019). This table reports the percent variance explained (PVE) by the linear dynamic eQTL model's fixed effects (excluding fixed effects related to cell line collapsed PCs) for all significant (eFDR <= .05) linear dynamic eQTLs. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For linear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, and then genotypeXday.

Hallmark gene set	0 PCs	1 PC	2 PCs	3 PCs	4 PCs	5 PCs
KRAS signalling dn	.0076	.0007	.472	1.0	1.0	1.0
Hypoxia	1	1	.33	.00095	.0048	.02
Myogenesis	.91	.01	1	.055	.011	.002
Interferon Gamma Response	1	.08	.39	.39	.086	.016

Hallmark gene set	6 PCs	7 PC	8 PCs	9 PCs	10 PCs
KRAS signalling dn	1.0	1.0	1.0	1.0	1.0
Hypoxia	.33	.022	1.0	1.0	.33
Myogenesis	.24	.055	.055	1.0	1.0
Interferon Gamma Response	.086	.0026	.086	1.0	.39

Table S2-6. Hallmark gene set enrichment of linear dynamic eQTLs. Bonferroni corrected p-values (Fisher's exact) from gene set enrichment within Hallmark gene sets (rows) of the 200 genes with the strongest linear dynamic eQTLs as a function of the number of cell line collapsed PCs used as covariates (columns). Only Hallmark gene sets with at least one significant enrichment (Bonferroni p-value <= .05) are shown.

Number of cell line collapsed PCs	Enrichment p-value
0	.08
1	.01
2	.01
3	.00099
4	6.8e-5
5	.00099
6	.01
7	.00099
8	.08
9	.08
10	.08

Table S2-7. Dilated cardiomyopathy gene set enrichment of linear dynamic eQTLs. p-values (Fisher's exact) from gene set enrichment within dilated cardiomyopathy gene set of the 200 genes with the strongest linear dynamic eQTLs as a function of the number of cell line collapsed PCs used as covariates.

Table S2-8. Percent variance explained for nonlinear dynamic eQTLs. Available as a text file online (Strober et al. 2019). This table reports the percent variance explained (PVE) by the nonlinear dynamic eQTL model's fixed effects (excluding fixed effects related to cell line collapsed PCs) for all significant (eFDR <= .05) nonlinear dynamic eQTLs. PVE for each covariate was estimated via ANOVA analysis which assumes an underlying order of covariates when iteratively computing the variance explained by each additional covariate. This was done to handle the covariance between covariates. For nonlinear dynamic eQTLs, covariates were ordered as follows: all cell line collapsed PC related terms, genotype, day, day^2, genotypeXday, and then genotypeXday^2.

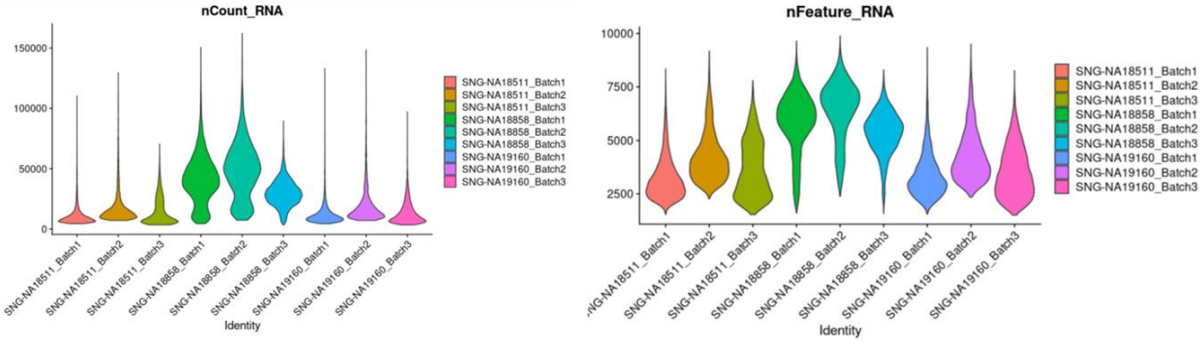


Fig. S3-1. UMI per cell and Genes per cell. Left: Violin Plot showing the total umi counts in cells from each individual in each replicate after filtering. Right: Violin Plot showing the number of genes (features) expressed in cells of each individual and each replicate after filter.

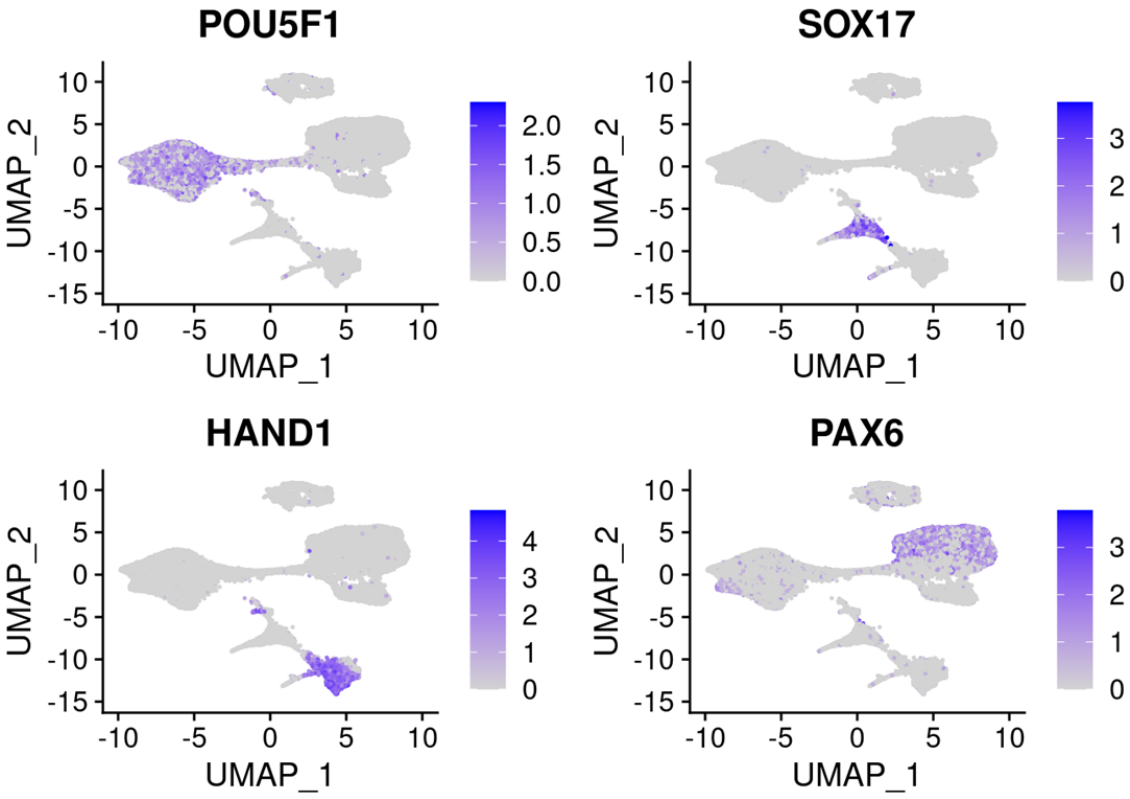


Fig. S3-2. Marker gene expression in EB cells. UMAP visualization of cells colored by expression of marker genes for pluripotent cells (POU5F1), endoderm (SOX17), mesoderm (HAND1), and ectoderm (PAX6). Color indicates the sum of the scaled and normalized counts.

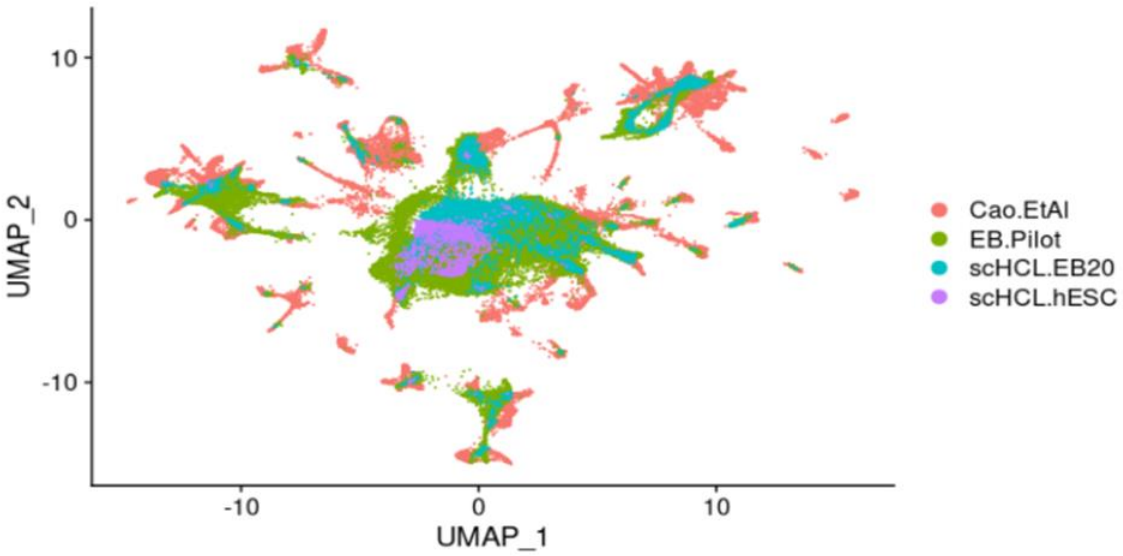


Fig. S3-3. UMAP visualization of EBs and integrated reference data sets. Points are colored by data set: cells from this study (EB.Pilot), cells from reference data sets of fetal cell types (Cao.EtAl), Externally generated Day 20 EBs (scHCL.EB20), and hESCs (hESC) after integration.

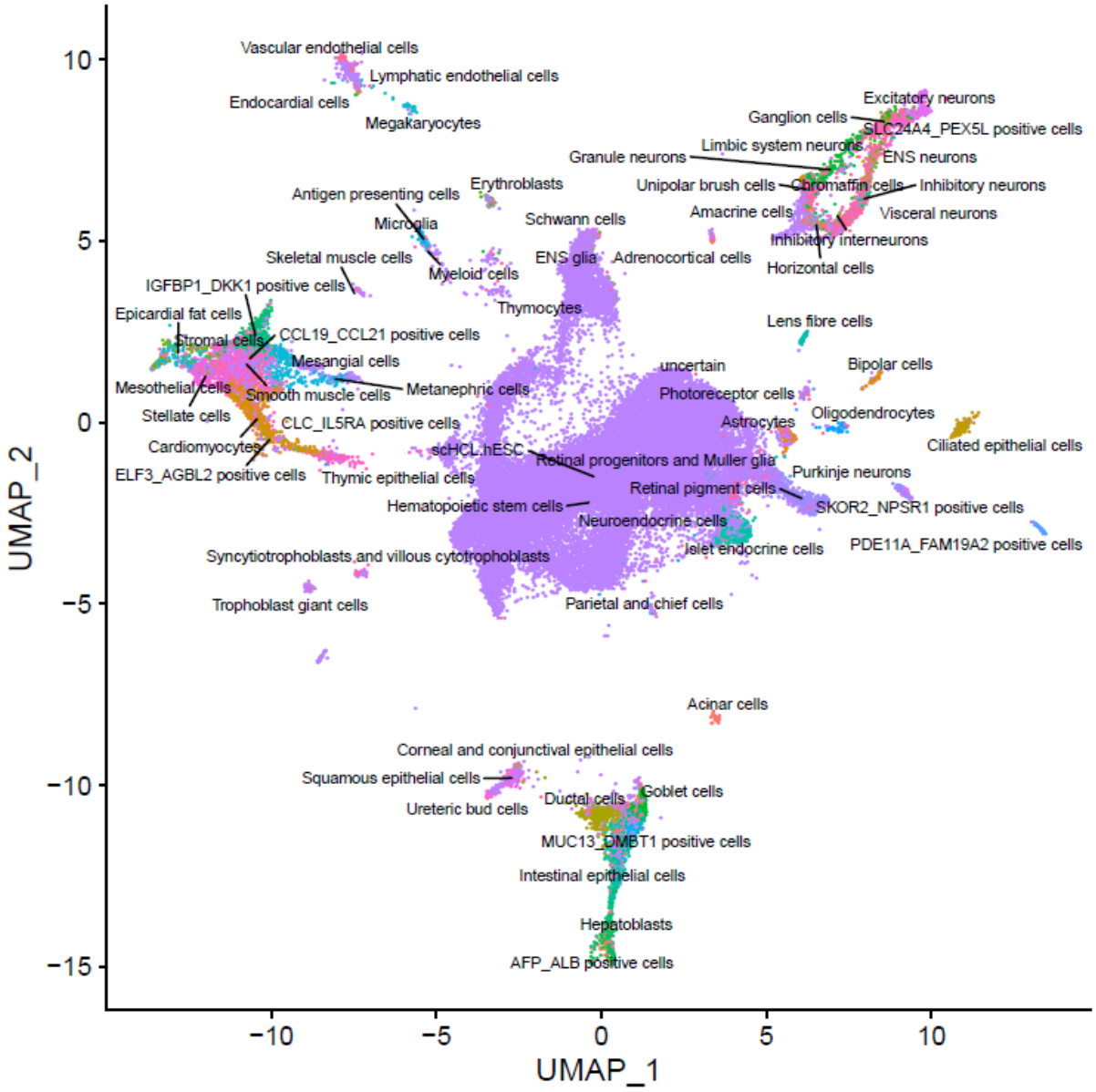


Fig. S3-4. UMAP visualization of EBs cells with annotations transferred from reference data. Points are colored by cell type annotation learned from the most similar reference cells. Cells which could not confidently be annotated are labelled as “uncertain”.

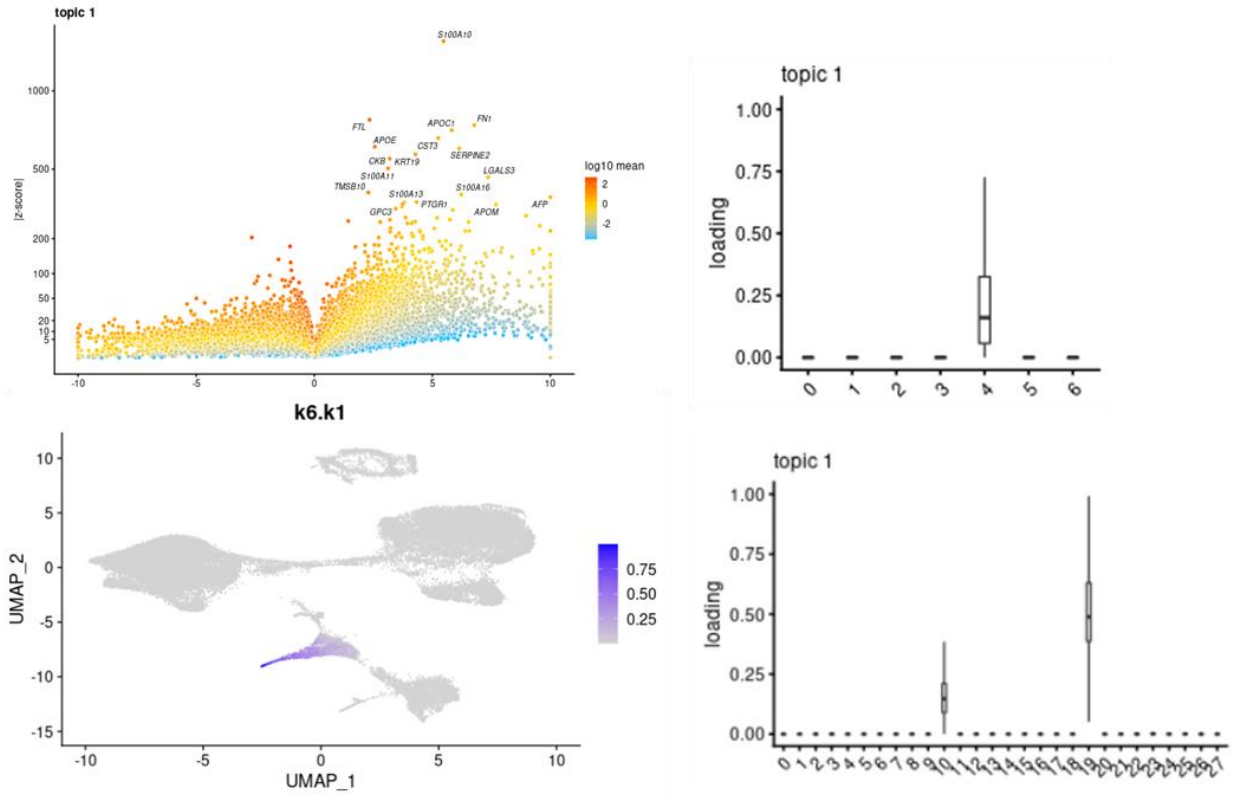


Fig. S3-5. Topic corresponds to hepatic cell fate. Top Left) Volcano Plot showing genes differentially expressed between topic 1 and all other topics from the k=6 topic analysis. Points are colored by the average count on the logarithmic scale. Bottom Left)) UMAP projection of cells colored by loading of topic 1 from k=6 topic analysis.Top Right) Bar plot showing the loading of topic 1 from the k=6 topic analysis on each seurat cluster at clustering resolution 0.1. Bottom Right)Bar plot showing the loading of topic 1 from the k=6 topic analysis on each seurat cluster at clustering resolution 1.

Topic	Top 10 driver genes
k1	S100A10, FTL, FN1, APOC1, CST3, SERPINE2, KRT19, CKB, S100A11, LGALS3, TMSB10, S100A16, AFP, PTGR1
k2	MT-CO2, MT-CO3, MT-CO1, MT-CYB, PRDX1, MT-ND4, MT-ATP6, GSTP1, MT-ND1, RPL8, APOE, RPSA, RPL12, PFN1, HMGA1
k3	PTMA, NCL, RPL23, SET, HSP90AB1, TPL27A, MT-ND4, L1TD1, SERBP1, TERF1, HSPD1, CENPF, DPPA4, MT-ATP6, UGP2
k4	S100A10, KRT19, S100A11, VIM, MDK, TMSB10, KRT8, SPARC, COL1A1, FN1, COL1A2, COL6A2, KRT18, TPM1, ANXA2
k5	TUBA1A, VIM, MARCKSL1, MARCKS, TUBA1B, MAP1B, ID3, CRABP1, PTMS, TMSB10, H1FX, STMN1, CENPV, CRABP2, NUCKS1
k6	RPS27, VIM, LDHA, GAPDH, IGFBP2, TUBA1A, RPL13, TMSB10, S100A10, RPL6, RPL30, RPL9, RPS19, RPL37

Table S3-1. Top driver genes of each topic from k=6 topic analysis. Top 10 driver genes of each topic based on Z-score from differential expression test.

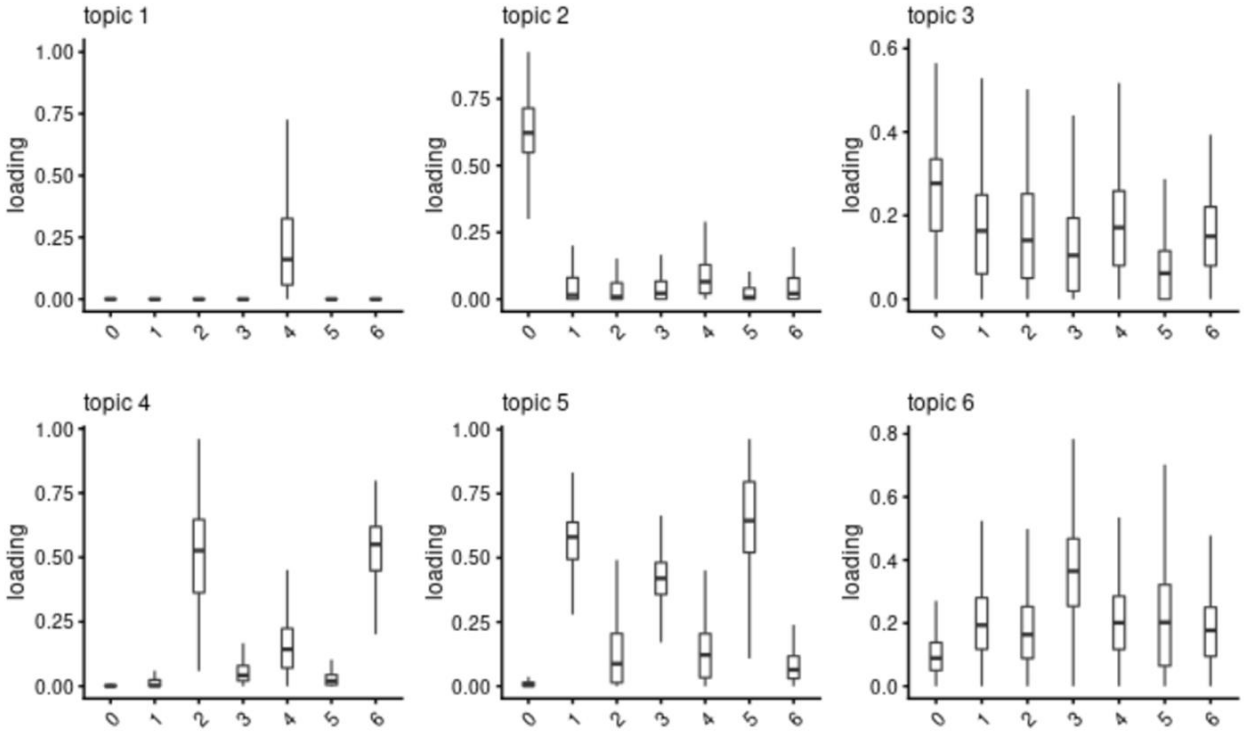


Fig. S3-6. Loading of topics on Seurat clusters. Barplots showing the loading of each topic (from the k=6 analysis) on each seurat cluster at resolution 0.1

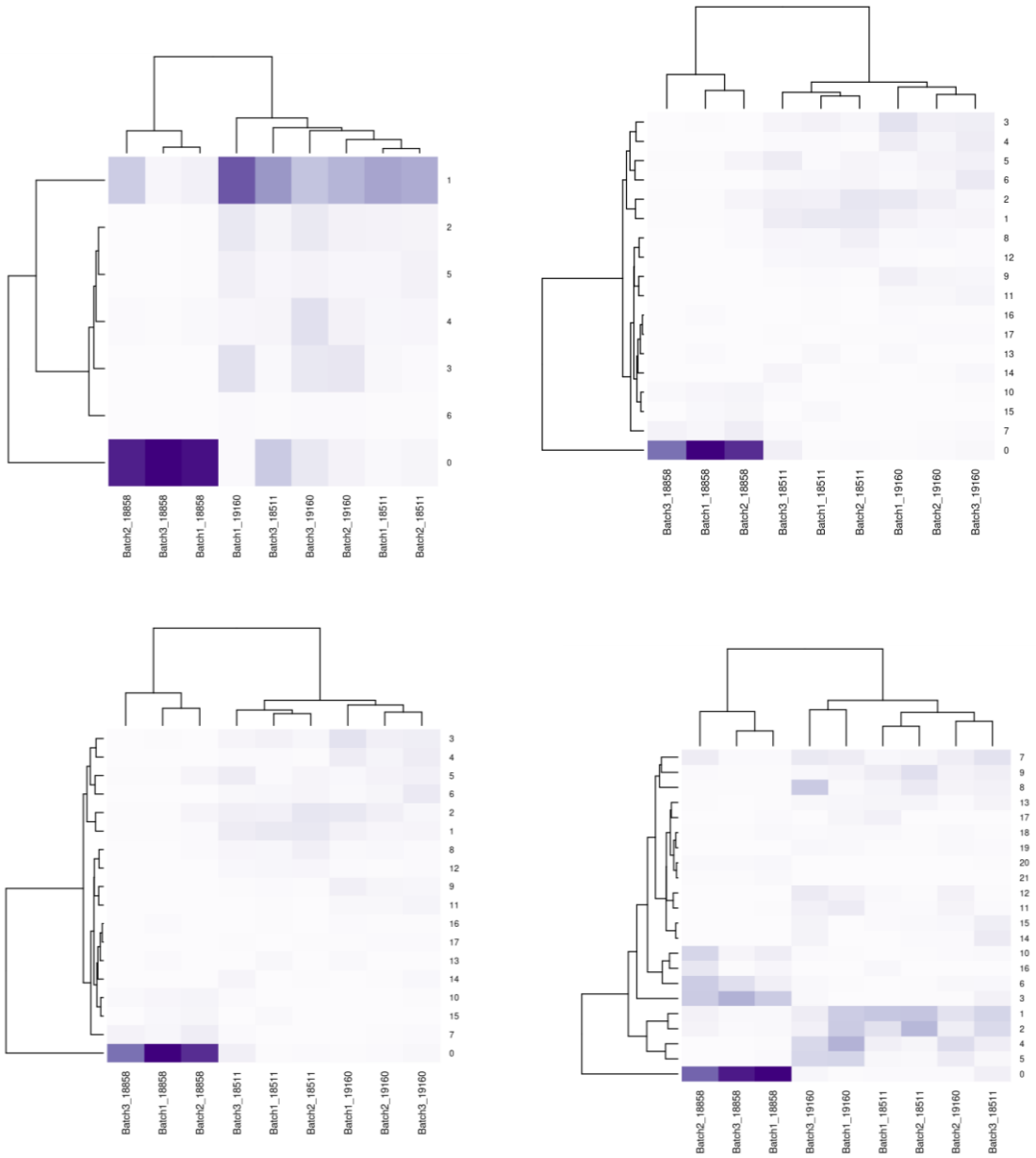


Fig. S3-7. Hierarchical clustering of samples based on cluster composition. Heatmaps showing hierarchical clustering of individual-replicate groups by the proportions of cells from each group assigned to each seurat cluster at resolution 0.5 (top left), resolution 0.5 (bottom left), resolution 0.8 (top right), and resolution 1 (bottom right).

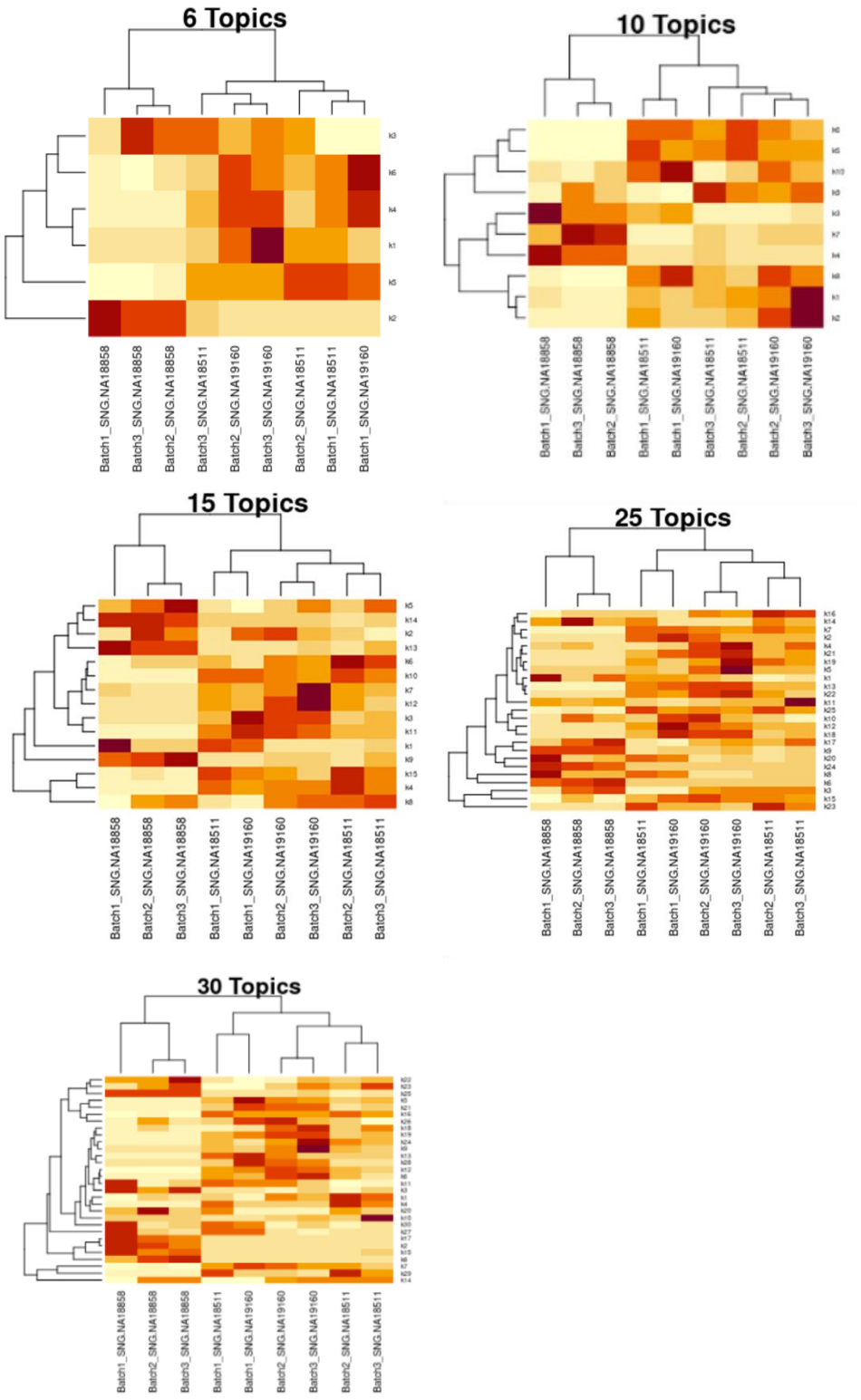


Fig. S3-8. Hierarchical clustering of samples based on topic loadings. Heatmaps showing hierarchical clustering of individual-replicate groups by the loading of each on topic on cells from topic models with varying values of k.

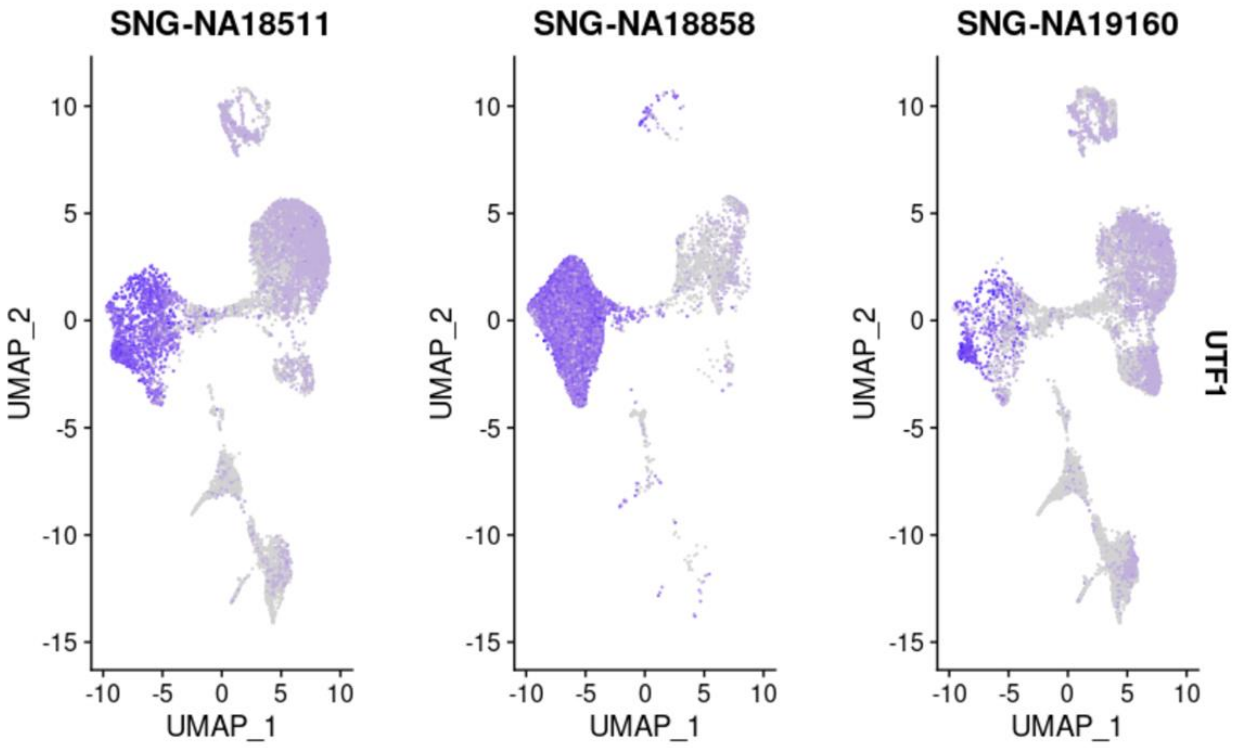


Fig. S3-9. Visualization of UTF1 expression in cells of each individual in UMAP space. Cells are colored by the sum of the scaled, normalized counts for UTF1.

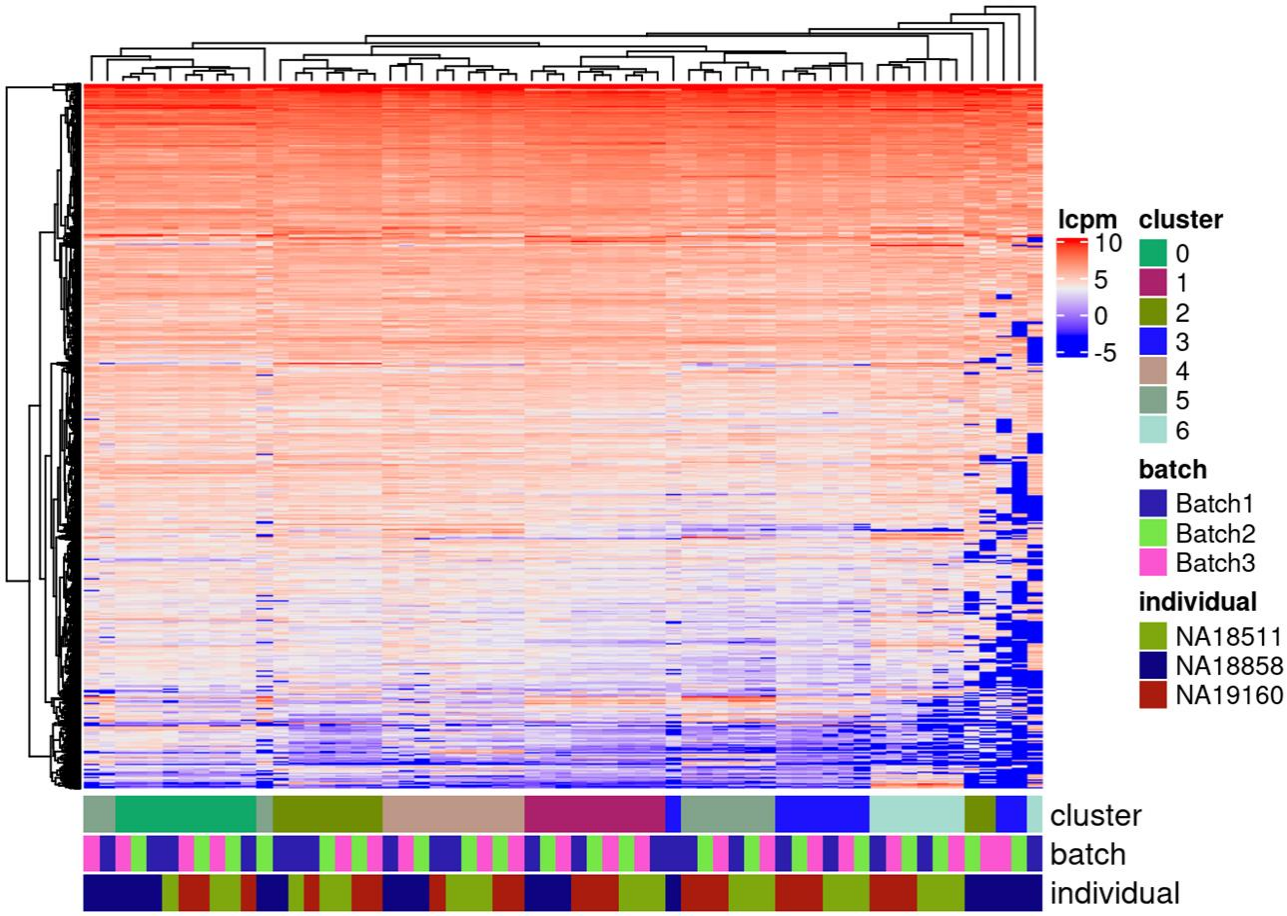


Fig. S3-10. Hierarchical clustering of samples based on gene expression. Heatmap showing hierarchical clustering of cells based on normalized gene expression. This analysis uses only gene expressed in at least 20% of cells in at least one cluster (at clustering resolution 0.1) and does not include ribosomal genes.

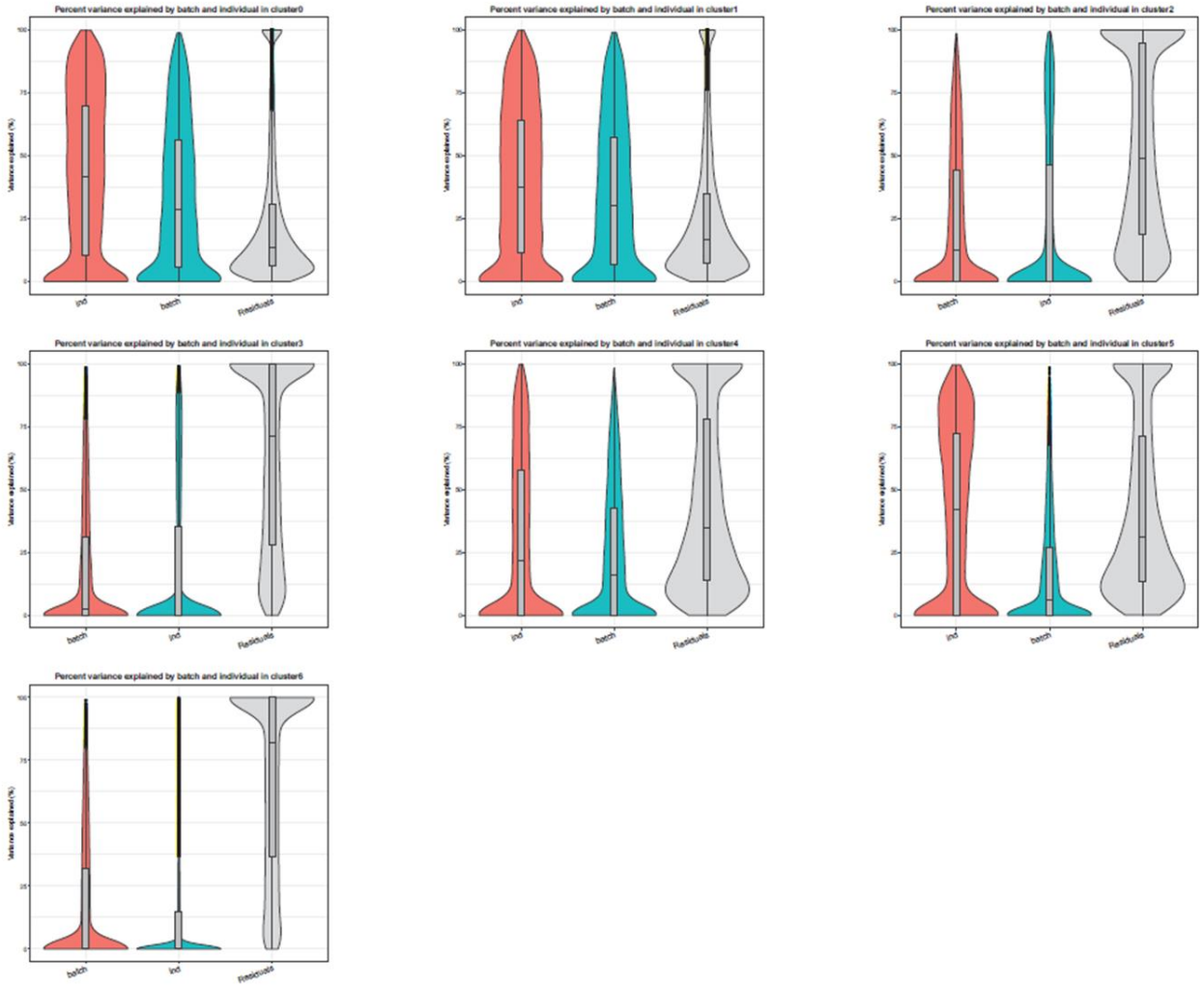
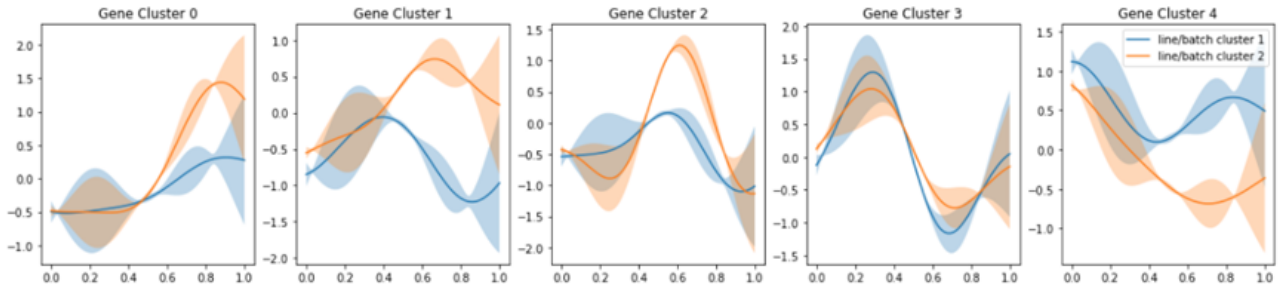


Fig. S3-11. Percent of gene expression variance explained by replicate and individual within cell type clusters. Violin plots showing the percent of variance in gene expression explained by replicate (batch) and individual (ind) in each seurat cluster (clustering resolution 0.1) partitioning variance in pseudobulk samples.

	0	1	2	3	4	5	6
18511_Batch1	99	2918	423	139	277	278	37
18511_Batch2	136	1958	35148	35	212	193	17
18511_Batch3	1508	2514	372	128	272	250	35
18858_Batch1	5995	526	29	28	116	35	3
18858_Batch2	4005	703	4	3	24	9	0
18858_Batch3	4882	235	5	9	47	12	0
19160_Batch1	81	2870	997	1129	196	634	63
19160_Batch2	179	1025	361	383	221	207	42
19160_Batch3	808	1634	747	819	1003	372	98

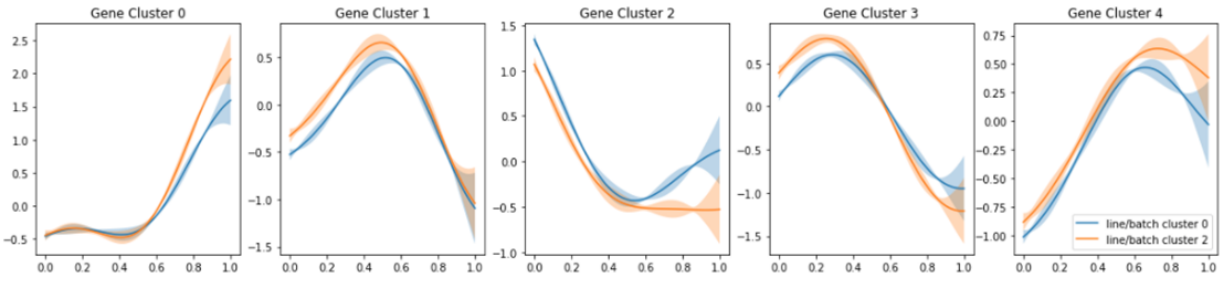
Table S3-2. Number of cells from each individual in each batch assigned to each Seurat cluster at clustering resolution 0.1.



	0	1	2	3	4
	bonferonni-adjusted	bonferonni-adjusted	bonferonni-adjusted	bonferonni-adjusted	bonferonni-adjusted
HALLMARK_TNFA_SIGNALING_VIA_NFKB	1.00714e-07	1	1	1	1
HALLMARK_G2M_CHECKPOINT	1	1	1	7.51983e-07	1
HALLMARK_APOPTOSIS	7.84836e-05	1	1	1	1
HALLMARK_ANDROGEN_RESPONSE	1	0.00706542	1	1	1
HALLMARK_APICAL_JUNCTION	0.000470291	1	1	1	1
HALLMARK_COMPLEMENT	1.37884e-05	1	1	1	1
HALLMARK_MTORC1_SIGNALING	1	1	1	1	6.57856e-06
HALLMARK_E2F_TARGETS	1	1	1	2.29089e-11	0.782356
HALLMARK_MYC_TARGETS_V2	1	1	1	1	0.00101437
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	1	1	1.43968e-05	1	1
HALLMARK_INFLAMMATORY_RESPONSE	0.00302993	1	1	1	1
HALLMARK_OXIDATIVE_PHOSPHORYLATION	1	1	1	1	0.00113012
HALLMARK_COAGULATION	0.000550209	1	1	1	1
HALLMARK_IL2_STAT5_SIGNALING	0.00380255	1	1	1	1
HALLMARK_KRAS_SIGNALING_UP	0.00372939	1	1	1	1

	cell line/batch	assignment
0	SNG-NA18511--Batch1	2
1	SNG-NA18511--Batch2	2
2	SNG-NA18511--Batch3	2
3	SNG-NA18858--Batch1	1
4	SNG-NA18858--Batch2	1
5	SNG-NA18858--Batch3	1
6	SNG-NA19160--Batch1	2
7	SNG-NA19160--Batch2	2
8	SNG-NA19160--Batch3	2

Fig. S3-12. Split-GPM clustering within the endothelial trajectory. Top) Table showing bonferonni-adjusted p-values from gene set enrichment analysis of gene modules. Bottom Left) Dynamic expression patterns of identified gene modules in each cluster of replicate-individual samples. Bottom Right) cluster assignments of each individual-batch sample based on shared patterns of dynamic gene expression



	0	2	3
	bonferonni-adjusted	bonferonni-adjusted	bonferonni-adjusted
HALLMARK_MITOTIC_SPINDLE	1	1	1.93314e-07
HALLMARK_G2M_CHECKPOINT	1	1	4.43539e-21
HALLMARK_ESTROGEN_RESPONSE_EARLY	1	0.000399656	1
HALLMARK_MTORC1_SIGNALING	1	0.00736101	4.21017e-05
HALLMARK_E2F_TARGETS	1	1	1.99314e-28
HALLMARK_MYC_TARGETS_V1	1	1	3.37513e-09
HALLMARK_MYC_TARGETS_V2	1	1	0.000866363
HALLMARK_OXIDATIVE_PHOSPHORYLATION	1	0.00171769	1
HALLMARK_GLYCOLYSIS	1	0.887663	0.00515952
HALLMARK_PANCREAS_BETA_CELLS	0.00843609	1	1

	cell line/batch	assignment
0	SNG-NA18511--Batch1	2
1	SNG-NA18511--Batch2	0
2	SNG-NA18511--Batch3	0
3	SNG-NA18858--Batch1	0
4	SNG-NA18858--Batch2	0
5	SNG-NA18858--Batch3	0
6	SNG-NA19160--Batch1	2
7	SNG-NA19160--Batch2	2
8	SNG-NA19160--Batch3	0

Fig. S3-13. Split-GPM clustering within the neural trajectory. Top) Table showing bonferonni-adjusted p-values from gene set enrichment analysis of gene modules. Bottom Left) Dynamic expression patterns of identified gene modules in each cluster of replicate-individual samples. Bottom Right) cluster assignments of each individual-batch sample based on shared patterns of dynamic gene expression

Supplementary Tables and Figures for Chapter IV

	19160	18511	18858	H28834	H21792	H28126	C40280	C3651	C3649
Batch 1	7932	5387	5398	5342	5873	6009	3958	5788	5454
Batch 2	4954	5179	4741	0	0	0	3685	4731	3620
Batch 3	7310	6201	5143	0	0	0	7302	4996	6207

Table S4-1. The number of high quality cells obtained from each individual in each replicate after filtering and quality control.

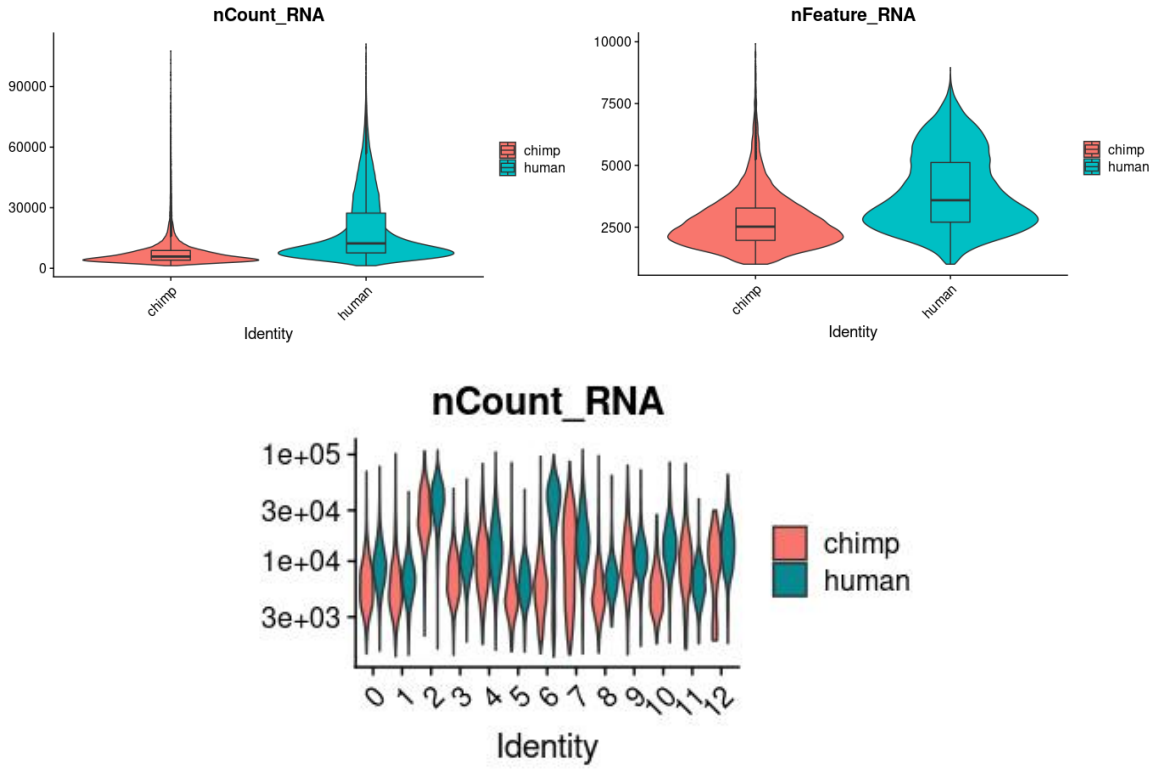


Fig. S4-1. Quality Control metrics after filtering. Top left) The number of UMI per cell from each species after filtering out low quality cells. Top right) The number of genes detected per cell from each species after filtering out low quality cells. The number of UMI per cell and genes per cell tends to be higher in humans, likely because the human sample includes more undifferentiated, pluripotent cells which have higher RNA content. Bottom) The number of UMI per cell in cells of each cluster (identified at clustering resolution 0.1) from each species. Log scaling of UMI counts is shown on the y axis.

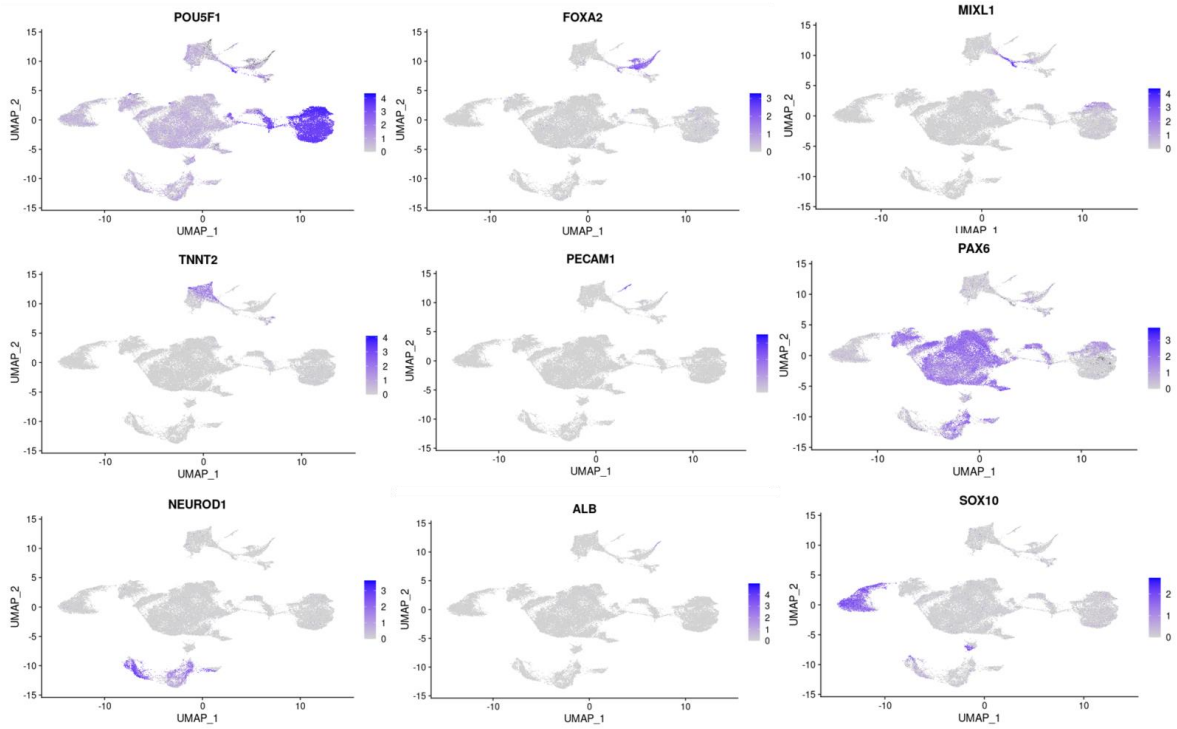


Fig. S4-2. Marker gene expression in UMAP space. Color indicates the sum of scale and normalized counts of known marker genes: POU5F1 (pluripotent), FOXA2 (definitive endoderm), MIXL1 (primitive streak), TNNT2 (cardiomyocytes, mesoderm), PECAM1 (endothelial cells), PAX6 (early ectoderm), NEUROD1 (neurons), ALB (hepatocytes), SOX10 (neural crest).

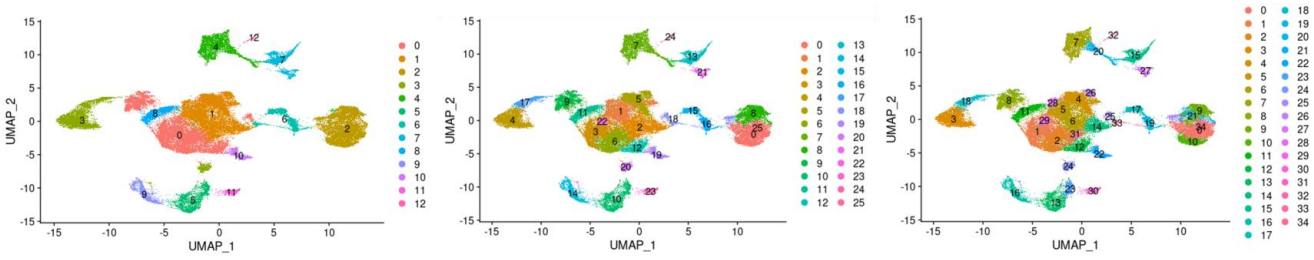


Fig. S4-3. UMAP visualization of clustering. Cells are colored by cluster assignment at clustering resolution 0.1 (left), 0.5 (center), 1 (right).

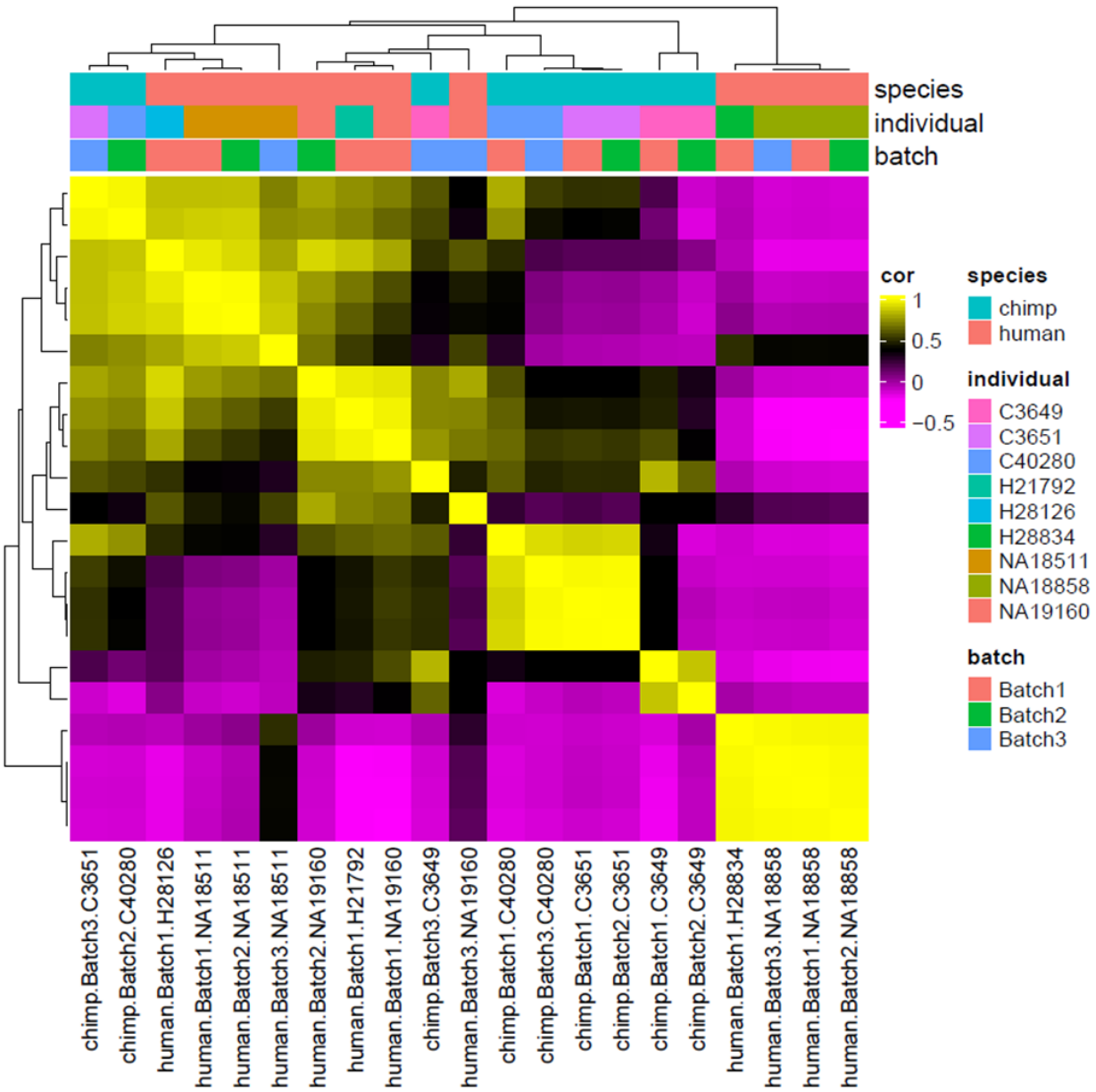


Fig. S4-4. Hierarchical clustering of samples based on cell type composition. hierarchical clustering of individual-replicate groups based on correlation of the proportions of cells assigned to each seurat cluster at resolution 0.1

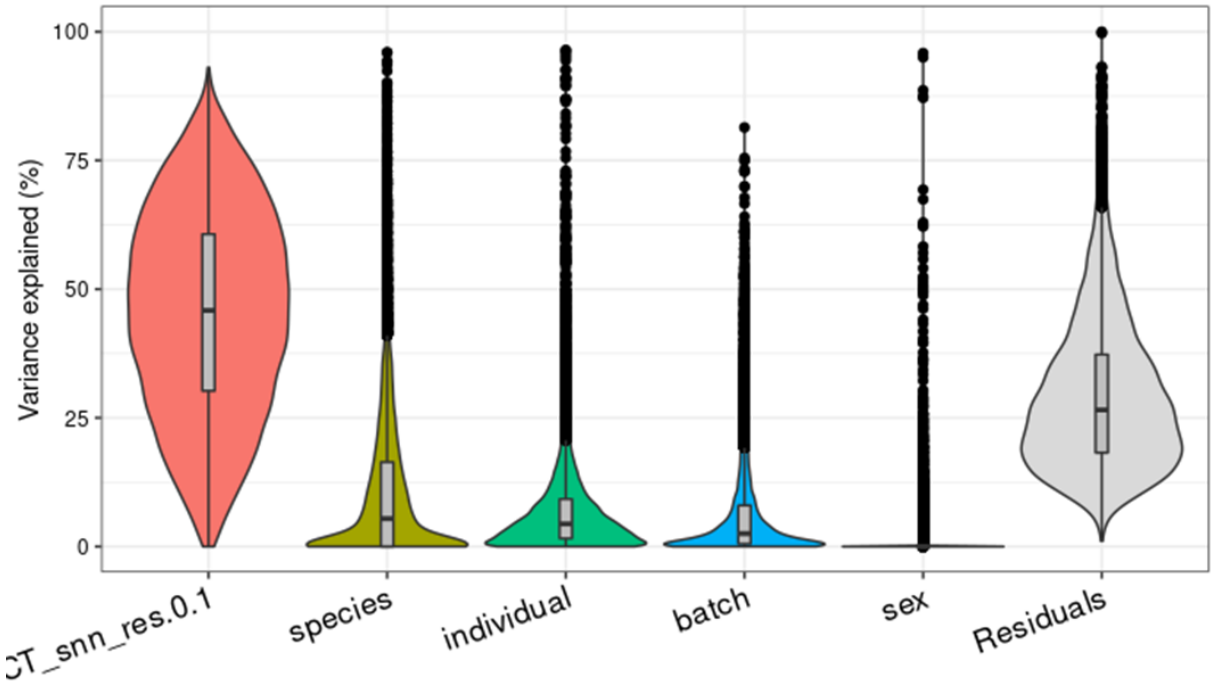


Fig. S4-5. Gene expression variance explained by biological and technical factors. Violin plot showing the percent of variance in gene expression explained by cluster (SCT_snn_res0.1, clusters defined at resolution 0.1), species, replicate (batch), sex, and individual in this data set after partitioning variance in pseudobulk samples.

Clustering resolution	% of genes significantly DE in at least 1 cluster
0.1	81.8
0.5	87.2
1	86.6

Table S4-2. Relationship of clustering resolution and number of DE genes. Table contains the percentage of tested genes that were significantly differentially expressed in at least one cluster at each clustering resolution.

individual	species	sex
NA19160	human	M
NA18511	human	F
NA18858	human	F
H28834	human	F
H21792	human	F
H28126	human	M
C40280	chimp	F
C3651	chimp	F
C3649	chimp	M

Table S4-3. iPSC line metadata.

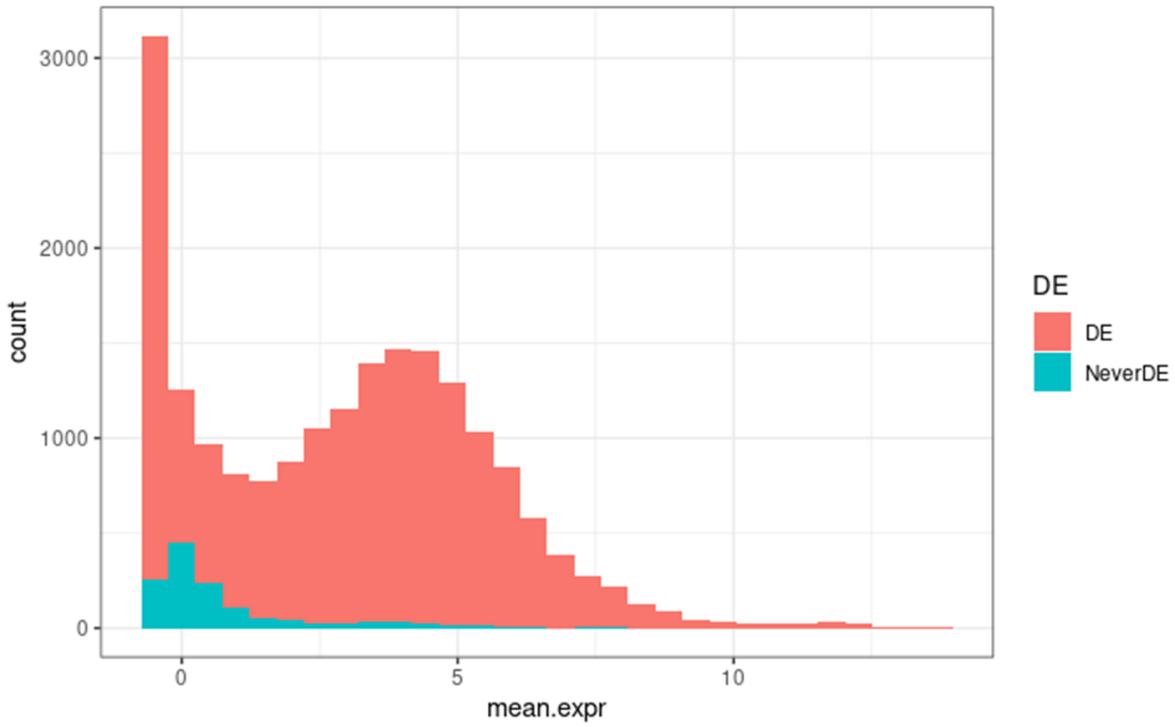


Fig. S4-6. Mean expression of DE genes. Mean expression of genes found to be significantly differentially expressed in at least 1 cluster at any clustering resolution (DE) compared to the mean expression of genes that were not DE in any cluster at any resolution (NeverDE).

GO.ID	Term	Annotated	Significant	Expected	result1
GO:0070988	demethylation	31	4	0.41	0.00067
GO:0070076	histone lysine demethylation	15	3	0.2	0.00089
GO:0016577	histone demethylation	16	3	0.21	0.00109
GO:0006353	DNA-templated transcription, termination	62	5	0.82	0.00126
GO:0043984	histone H4-K16 acetylation	17	3	0.22	0.00131
GO:0006482	protein demethylation	18	3	0.24	0.00156
GO:0008214	protein dealkylation	18	3	0.24	0.00156
GO:0044728	DNA methylation or demethylation	41	4	0.54	0.00196
GO:0016569	covalent chromatin modification	304	11	4	0.00201
GO:0006397	mRNA processing	403	13	5.3	0.00219
GO:0016071	mRNA metabolic process	670	18	8.81	0.00241
GO:0051276	chromosome organization	755	19	9.93	0.00378
GO:0006363	termination of RNA polymerase I transcri...	27	3	0.36	0.00513
GO:0042752	regulation of circadian rhythm	54	4	0.71	0.00537
GO:0006304	DNA modification	55	4	0.72	0.00573

Table S4-4. Gene Ontology enrichments of genes with conserved expression patterns.
The top 15 most significant GO term enrichments of genes with a conserved pattern of gene expression (the top 200 most highly expressed genes that were never significantly differentially expressed in any cluster at any clustering resolution).

GO.ID	Term	Annotated	Significant	Expected	result1
GO:0032354	response to follicle-stimulating hormone	11	2	0.03	0.00034
GO:0034698	response to gonadotropin	14	2	0.04	0.00056
GO:0042177	negative regulation of protein catabolic...	103	3	0.26	0.00219
GO:0071108	protein K48-linked deubiquitination	29	2	0.07	0.00242
GO:0046686	response to cadmium ion	40	2	0.1	0.00458
GO:0030162	regulation of proteolysis	456	5	1.16	0.00524
GO:1901799	negative regulation of proteasomal prote...	49	2	0.12	0.00681
GO:1901655	cellular response to ketone	55	2	0.14	0.00852
GO:1903051	negative regulation of proteolysis invol...	63	2	0.16	0.01106
GO:0050810	regulation of steroid biosynthetic proce...	64	2	0.16	0.0114
GO:0045861	negative regulation of proteolysis	187	3	0.48	0.01152
GO:1903363	negative regulation of cellular protein ...	73	2	0.19	0.01465
GO:0072330	monocarboxylic acid biosynthetic process	206	3	0.52	0.01493
GO:0019218	regulation of steroid metabolic process	82	2	0.21	0.01826
GO:0009895	negative regulation of catabolic process	225	3	0.57	0.01887

Table S4-5. Gene Ontology enrichments of mesoderm-specific DE genes. The top 15 most significant GO term enrichments of cell type-specific genes identified in cluster 4 at clustering resolution 0.1 (cluster represents mesodermal cells based on marker genes expression).

GO.ID	Term	Annotated	Significant	Expected	result1
GO:1905515	non-motile cilium assembly	47	5	0.49	0.00012
GO:0001662	behavioral fear response	16	3	0.17	0.00057
GO:0042596	fear response	16	3	0.17	0.00057
GO:0002209	behavioral defense response	17	3	0.18	0.00068
GO:0061512	protein localization to cilium	43	4	0.45	0.00103
GO:1902017	regulation of cilium assembly	44	4	0.46	0.00112
GO:0060271	cilium assembly	281	9	2.94	0.0027
GO:0033555	multicellular organismal response to str...	27	3	0.28	0.00272
GO:0044782	cilium organization	289	9	3.03	0.00326
GO:0048666	neuron development	726	16	7.6	0.00352
GO:0120031	plasma membrane bounded cell projection ...	414	11	4.34	0.00398
GO:0002576	platelet degranulation	64	4	0.67	0.00446
GO:0030031	cell projection assembly	421	11	4.41	0.00451
GO:1902855	regulation of non-motile cilium assembly	10	2	0.1	0.00463
GO:0120036	plasma membrane bounded cell projection ...	1038	20	10.87	0.00512

Table S4-6. Gene Ontology enrichments of neuron-specific DE genes. The top 15 most significant GO term enrichments of cell type-specific genes identified in cluster 5 at clustering resolution 0.1 (cluster represents neuronal cells based on marker genes expression).

GO.ID	Term	Annotated	Significant	Expected	result1
GO:0019827	stem cell population maintenance	125	21	8.6	0.00011
GO:0098727	maintenance of cell number	126	21	8.67	0.00013
GO:0070125	mitochondrial translational elongation	85	16	5.85	0.00019
GO:0070126	mitochondrial translational termination	86	16	5.92	0.00022
GO:0006415	translational termination	101	17	6.95	0.00049
GO:0044380	protein localization to cytoskeleton	52	11	3.58	0.00068
GO:0071539	protein localization to centrosome	30	8	2.06	0.00073
GO:0010467	gene expression	3879	307	266.95	0.00078
GO:1905508	protein localization to microtubule orga...	31	8	2.13	0.00092
GO:0006112	energy reserve metabolic process	56	11	3.85	0.0013
GO:0072698	protein localization to microtubule cyto...	48	10	3.3	0.00134
GO:0034341	response to interferon-gamma	93	15	6.4	0.0016
GO:0044264	cellular polysaccharide metabolic proces...	66	12	4.54	0.00162
GO:0090110	cargo loading into COPII-coated vesicle	14	5	0.96	0.00181
GO:0019221	cytokine-mediated signaling pathway	396	43	27.25	0.00185

Table S4-7. Gene Ontology enrichments of early developmental cell type-specific DE genes. The top 15 most significant GO term enrichments of cell type-specific genes identified in cluster 6 at clustering resolution 0.1 (cluster represents early developmental cells transitioning from pluripotency to ectodermal lineage based on marker genes expression).

GO.ID	Term	Annotated	Significant	Expected	result1
GO:0032456	endocytic recycling	39	2	0.06	0.0019
GO:0006892	post-Golgi vesicle-mediated transport	98	2	0.16	0.0113
GO:0045165	cell fate commitment	118	2	0.2	0.0161
GO:0060968	regulation of gene silencing	118	2	0.2	0.0161
GO:0016584	nucleosome positioning	11	1	0.02	0.0182
GO:0021895	cerebral cortex neuron differentiation	12	1	0.02	0.0198
GO:0072148	epithelial cell fate commitment	12	1	0.02	0.0198
GO:0002115	store-operated calcium entry	13	1	0.02	0.0214
GO:0006878	cellular copper ion homeostasis	13	1	0.02	0.0214
GO:0008209	androgen metabolic process	13	1	0.02	0.0214
GO:0031936	negative regulation of chromatin silenci...	13	1	0.02	0.0214
GO:0030206	chondroitin sulfate biosynthetic process	15	1	0.02	0.0247
GO:0032011	ARF protein signal transduction	15	1	0.02	0.0247
GO:0032012	regulation of ARF protein signal transdu...	15	1	0.02	0.0247
GO:0008210	estrogen metabolic process	16	1	0.03	0.0263

Table S4-8. Gene Ontology enrichments of endoderm-specific DE genes. The top 15 most significant GO term enrichments of cell type-specific genes identified in cluster 7 at clustering resolution 0.1 (cluster respresents endodermal cells based on marker genes expression).