THE UNIVERSITY OF CHICAGO

INVESTIGATING THE DYNAMICS OF GENE REGULATION DURING CARDIOMYOCYTE DIFFERENTIATION USING BULK AND SINGLE CELL RNA SEQUENCING

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

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM: GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

REEM ELORBANY

CHICAGO, ILLINOIS

DECEMBER 2020

This dissertation is dedicated to my parents, Ola Mobasher and Mohammad Elorbany.

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

I am grateful to my advisor, Dr. Yoav Gilad, for his guidance and support in helping me grow as a scientist. I would also like to thank everyone in the Gilad lab, all of whom enriched my time as a graduate student: Katherine Rhodes, Briana Mittleman, Lauren Blake, Ittai Eres, Michelle Ward, Kenneth Barr, Genevieve Housman, Anthony Hung, Wenhe Lin, Deji Adegunsoye, Ben Umans, Sebastian Pott, Joyce Hsiao, Po Yuan Tung, Benjamin Fair, Claudia Garcia, Jonathan Burnett, Emilie Briscoe, Natalia Gonzales, and John Blischak.

I would like to thank my collaborators and those who supported my work in other labs, including: Benjamin Strober, Nirmal Krishnan, Karl Tayeb, and Josh Popp of Dr. Alexis Battle's lab at Johns Hopkins; and Heather Eckart, Ryan Dohn, and Rebecca Back of Dr. Anindita Basu's lab at the University of Chicago.

I would also like to thank my committee members, Drs. Anindita Basu, Mengjie Chen, and Marcus Clark, for their feedback, advice, and guidance over the development of my projects.

I am grateful to my classmates and friends in the Medical Scientist Training Program, the Department of Human Genetics, and the Committee on Genetics, Genomics, and Systems Biology.

This work would not have been possible without the endless support of my parents, Ola Mobasher and Mohammad Elorbany, and my sister Heba.

Abstract

Genetic variants that alter gene regulation play a crucial role in the genetics of human development and disease. Although genome-wide association studies have found thousands of genetic loci associated with complex phenotypes, a substantial fraction of these trait-associated loci remain unexplained. Gene regulatory effects are context-specific and can result in dynamic gene expression changes over time and across cell types. To enhance our understanding of the genetic architecture of complex traits in a dynamic context, we studied dynamic genetic regulation of gene expression during cardiomyocyte differentiation using bulk and single-cell RNA-sequencing. We generated time-series gene expression data over multiple differentiation time points and mapped expression quantitative trait loci (eQTLs), or variants whose effects on expression are modulated by differentiation time. We identified hundreds of dynamic eQTLs which change over time, including nonlinear eQTLs which affect only intermediate stages of differentiation. Using single-cell RNA-sequencing, we were able to disentangle the effects of gene regulatory variation from variation in cell type composition, both of which may change over a differentiation time course under the influence of genetic factors. Together, these studies demonstrate the use of time series data to investigate the dynamics of gene regulation and cell type composition changes over time, and provide new insight into the genetic architecture underlying human development, complex phenotypes, and disease.

Chapter I: Introduction

Gene regulation in genomic studies

A primary aim of human genetics and genomics research is to understand the genetic architecture of complex phenotypes and human diseases. Although genetic factors are known to play a role in complex phenotypes such as cardiovascular disease, much of the specific genetic contribution to these phenotypes remains unknown (Bis et al. 2011, Myocardial Infarction Genetics Consortium 2009, Manolio et al. 2009, Eichler et al. 2010). A greater understanding of the genetic causes or associated molecular phenotypes of complex diseases like heart disease could enable the use of personalized medicine to identify an individual's risk based on these factors, or potentially to develop novel techniques for treatment or prevention (van der Wijst et al. 2018).

We can identify regions of the genome associated with a particular phenotype or disease using genome-wide association studies, or GWAS. Previous genome-wide association studies have found many genomic loci associated with increased risk of cardiovascular disease or heart failure (Wellcome Trust Case Control Consortium 2007, CARDIoGRAMplusC4D Consortium et al. 2013, Shah et al. 2020). In many cases, the function of these loci is unclear, as are the contexts in which they are expressed, and their potential downstream consequences. One potential downstream mechanism of a disease-associated gene might be translation into a functional protein, which is then involved in some disease pathway. However, the majority of trait-associated genomic loci identified by GWAS are located in non-protein coding regions of the genome (Edwards et al. 2013). Rather than affecting protein translation, non-coding loci are thought to be involved in regulating when, where, and how much a gene is expressed (Britten and Davidson 1969). Variation in a regulatory region between individuals may lead to inter-individual differences in gene regulation, as measured by gene expression levels of an associated coding region or other molecular phenotypes. To gain a more complete understanding of the genetic architecture of human phenotypes and disease, we must study how genetic variation contributes to gene regulation.

Studying gene regulation in a dynamic context

One way to study gene regulation is by performing quantitative trait locus, or QTL, analysis. A quantitative trait locus is a location in the genome where a person's genotype at that locus is associated with variation in a quantitative trait, such as gene expression, chromatin accessibility, or DNA methylation levels. QTLs specifically focused on gene expression (expression QTLs, or eQTLs) have been identified in a wide variety of cell types and contexts;

their study has greatly contributed to the understanding of gene regulation (GTEx Consortium 2017; Lappalainen et al. 2013; Battle et al. 2014; Pickrell et al. 2010; Stranger et al. 2012), and in many cases, eQTLs have enhanced our understanding of disease (Nica et al. 2010; Nicolae et al. 2010). For example, analyzing the results of an eQTL study in conjunction with GWAS results from a relevant trait can reveal how genetic regulation of gene expression influences the development of that trait (Pickrell 2014).

In a typical eQTL study, gene expression is assayed in a particular tissue or cell type, in a static environment and at a static point in developmental time. Thus, a typical eQTL study can provide information on gene regulation only in the particular context in which the assay was performed. However, gene regulation depends on cell type and environment, which can change over time. For example, the gene regulatory profile of an undifferentiated pluripotent cell can be quite different from that of a fully differentiated heart muscle cell (GTEx Consortium 2017). The typical eQTL study does not provide an opportunity to investigate the dynamics of gene regulation, or how gene regulatory effects can change over time. 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 cardiomyocytes as a model system

To study the dynamics of gene regulation in a biologically relevant model, we can focus on a differentiation time course in which a cell transitions from one cell type to another, such as the differentiation of pluripotent cells in early human development. During development, a single pluripotent cell proliferates and differentiates into the multitude of diverse cell types that exist in

the adult human body (Alberts et al. 1989). Each of those terminal cell types has its own unique gene expression profile, which can change over time as the cell differentiates (GTEx Consortium 2017). By studying the dynamics of gene regulation in this context, we can learn how the genome directs when, where, and how different cell types arise during development.

A developmental or cell type differentiation time course can be studied using induced pluripotent stem cells and their derived terminal cell types as a model system. Induced pluripotent stem cells, or iPSCs, are cells that have been reprogrammed from adult somatic cells into an induced state of self-renewal and pluripotency (Yu et al. 2007; Okita et al. 2011). iPSCs are an appealing model system because of their capacity to undergo directed differentiation into theoretically any somatic cell type in the human body. This property of iPSCs can allow us to study not only the starting point (undifferentiated cell) and end point (differentiated cell), but also intermediate steps in between, and the dynamic process of differentiation overall. Unlike using donor tissue samples, using iPSCs enables us to investigate gene expression at many intermediate time points using the exact same genotypes as samples taken from the undifferentiated and terminal cell types.

This project aims to study the dynamics of gene regulation during cardiac development, specifically in cardiomyocytes, or heart muscle cells. Cardiac cell types are of special interest due to the high prevalence and severity of pathologic cardiac phenotypes and disease -- cardiovascular disease is the leading cause of death in the United States (Heron 2019). Early human heart development is a complex process, which involves undifferentiated cells becoming terminal cell types such as cardiomyocytes, cardiac fibroblasts, and endocardial cells, and consolidating into the final 3-dimensional structure of the heart (Brade et al. 2013). Any cell type in the human heart (including cardiomyocyte, fibroblast, endothelial cell, and vascular smooth

muscle cell) is potentially relevant to our understanding of cardiac pathology (Nakano et al. 2012). However, the ultimate outcome of healthy heart function is the ability to pump blood, which is based on cardiomyocyte performance. Numerous studies have used iPSC- derived cardiomyocytes to gain insight into cardiovascular diseases (Oh et al. 2012; Narsinh, Narsinh, and Wu 2011; Zhi et al. 2012; Yazawa et al. 2011; Lu and Yang 2011; Josowitz et al. 2011; Dambrot et al. 2011). Analysis of gene regulation in cardiomyocytes can be combined with information on cardiovascular disease from published genome-wide association studies to identify genetic variants that influence disease risk. By using a cardiomyocyte developmental time course, we can study the dynamics of gene regulation, and ultimately learn more about cardiac phenotypes and cardiac-related pathologies.

Induced pluripotent stem cells can be differentiated into iPSC-derived cardiomyocytes using a relatively straightforward protocol, which involves first activating and then inhibiting the Wnt pathway (which controls heart differentiation in vivo) (Lian et al. 2013; Burridge et al. 2014). iPSC-derived cardiomyocytes form visible 3-dimensional structures and exhibit synchronization and mechanical beating, similar to primary cardiomyocytes. The differentiated cells also show an increased level of expression of cardiac marker genes, including cardiac troponin T, a cell-specific component of cardiac muscle. Previous studies have shown that gene expression data from iPSC-derived cardiomyocytes are more similar to gene expression data from post-mortem heart tissue samples than any other tissue sample from the Genotype-Tissue Expression (GTEx) project (Banovich et al. 2018). Similarly, eQTLs identified in iPSC-derived cardiomyocytes are most enriched with eQTLs identified in primary heart tissues compared to eQTLs identified in any other GTEx tissues (Banovich et al. 2018). These findings indicate that

iPSC-derived cardiomyocytes broadly recapitulate the gene expression profile that exists in the human heart, and can be a useful model to study cardiac genetics.

It is important to note that in vitro differentiations will typically not produce a cell culture with 100% purity of the intended terminal cell type, and purity may vary by individual cell line. Despite these limitations, the capacity for self-renewal and controlled differentiation make iPSCs a highly useful system for studying gene regulation during cellular development.

Bulk and single-cell RNA sequencing

Traditionally, experiments studying gene expression levels have been limited to bulk RNA sequencing technologies, which provide an average gene expression measurement of all cells in the sample (Trapnell 2015). Thus, bulk RNA-sequencing technology can lose valuable information, especially in a differentiation time course. During a differentiation, the cell types within a sample could be changing day-by-day, from undifferentiated iPSC, to early mesoderm, to cardiac precursor, to cardiomyocyte, for example. The proportions of these different cell types within a sample could change both over time and between different individuals, whose differentiation may progress at different rates or exhibit different trajectories. Additionally, as previously noted, an *in vitro* differentiation does not result in a terminal cell culture with 100% purity; other cell types may be present as a result of failed or off-target differentiations.

Single-cell RNA sequencing can provide new insight into gene expression changes at deeper resolution and with more nuance than bulk RNA sequencing. Instead of averaging gene expression from all cells in a sample, single-cell sequencing can be used to investigate cell type composition, as it provides a unique gene expression measurement for each individual cell within

a sample (Trapnell 2015; Macosko et al. 2015). For experiments involving differentiation, single-cell RNA-sequencing enables the characterization of cell-to-cell variation and composition without the need to purify populations of interest (Trapnell 2015; Treutlein et al. 2014). Sequencing single cells also allows us to investigate overall differences in differentiation trajectories between individuals. We can use single-cell RNA-sequencing to ask whether gene expression difference between samples is due to a true gene regulatory difference, or due to cell composition difference, or both. This is especially pertinent to a differentiation time course study, as our dynamic gene expression measurements can be affected by both gene regulatory changes and cell composition changes over time. Single-cell sequencing technology provides the opportunity to investigate individual variation in gene regulation on a cellular level, to observe differentiation trajectories at higher resolution, and to discern the effects of changing cell type proportions in a sample over time.

Summary

In this thesis, I investigate the dynamics of gene regulation during human iPSC to cardiomyocyte differentiation. First, I use bulk RNA-sequencing in a high-resolution time course to identify locations in the genome where natural variation between individuals is associated with differences in gene expression which are modulated by differentiation time. Then, I use single-cell RNA-sequencing with the same individual cell lines to distinguish the effects of gene regulation from the effects of cell type composition changes over time. Dynamic gene regulatory loci identified throughout the differentiation time course can provide insight into transient effects not found in mature tissues, and provide a resource for investigating mechanisms underlying previously unexplained associations with disease and other phenotypes.

Chapter II: Dynamic genetic regulation of gene expression during cardiomyocyte differentiation using bulk RNA-seq

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.

[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.

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 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 (Yi et al. 2106, Albert et al. 2015). 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 (Albert et al. 2015). Expression quantitative trait locus (eQTL) analysis has shown that many disease-associated loci influence the regulation of nearby genes (Zhu et al. 2016, Nicolae et al. 2010), but a substantial fraction of disease-associated loci still remain unexplained (Joehanes et al. 2017, Wen et al. 2017).

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 (GTEx Consortium 2017, Knowles et al. 2017) 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 (GTEx Consortium 2017), 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 (Banovich et al. 2018). 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) (Banovich et al. 2018, *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 (Okita et al. 2007, Lian et al. 2012) (fig. S2-5). As expected (Lian et al. 2012), cardiomyocyte purity and the expression of lineage marker genes are variable across our samples. This variability between cell lines was observed 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).

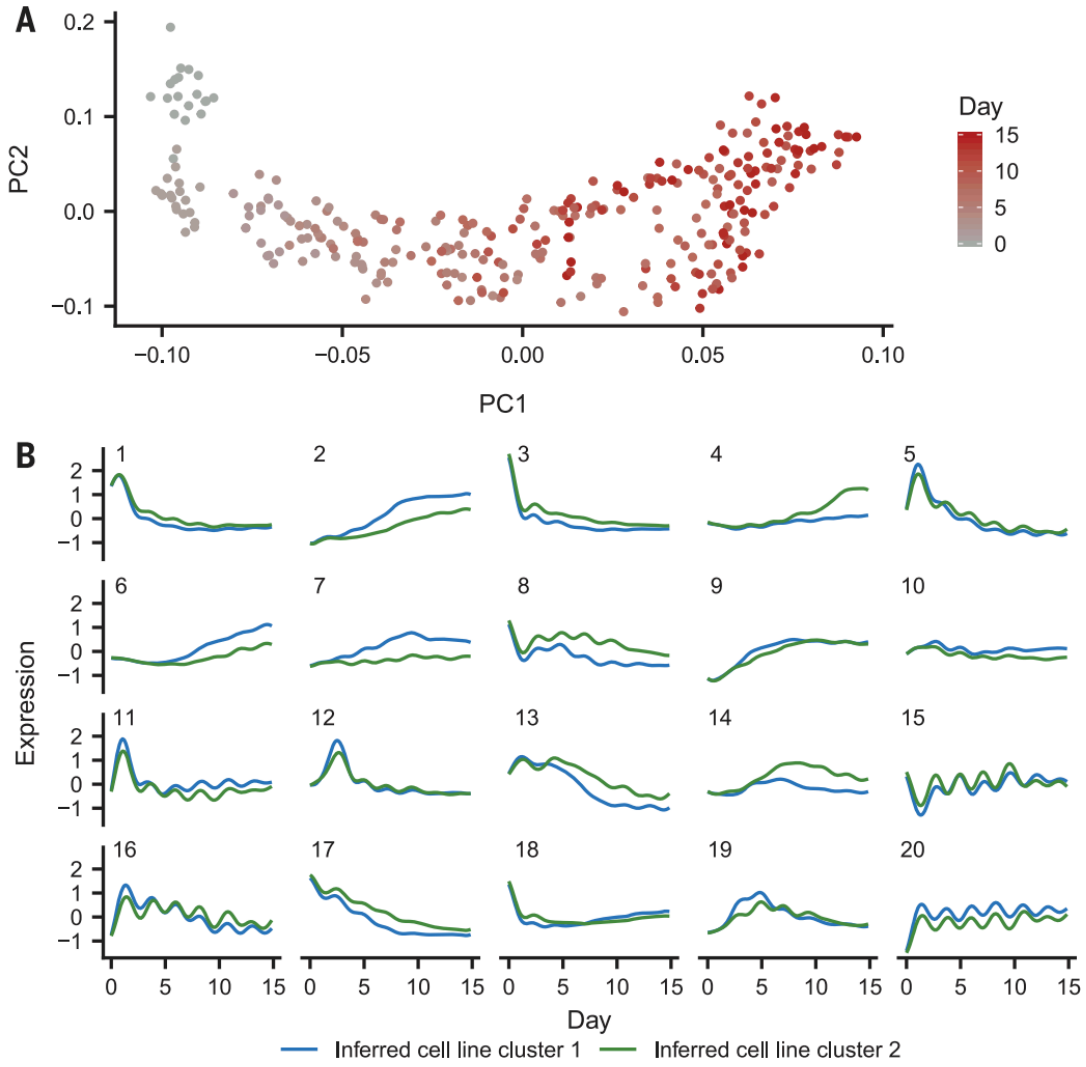


Fig. 2-1. Gene expression trends throughout cardiomyocyte differentiation. (A) The first two gene expression principal component (PC) loadings for all 297 RNA-seq samples across cell lines, where each sample is colored according to the day of collection. (B) Predicted cell line cluster expression trajectories 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.

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 × 10−14) (gene cluster 2 in Fig. 2-1B) (Liberzon et al. 2015). 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 (Kubara et al. 2018). 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 (van de Geijn et al. 2015) 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 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) (Banovich et al. 2018).

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.

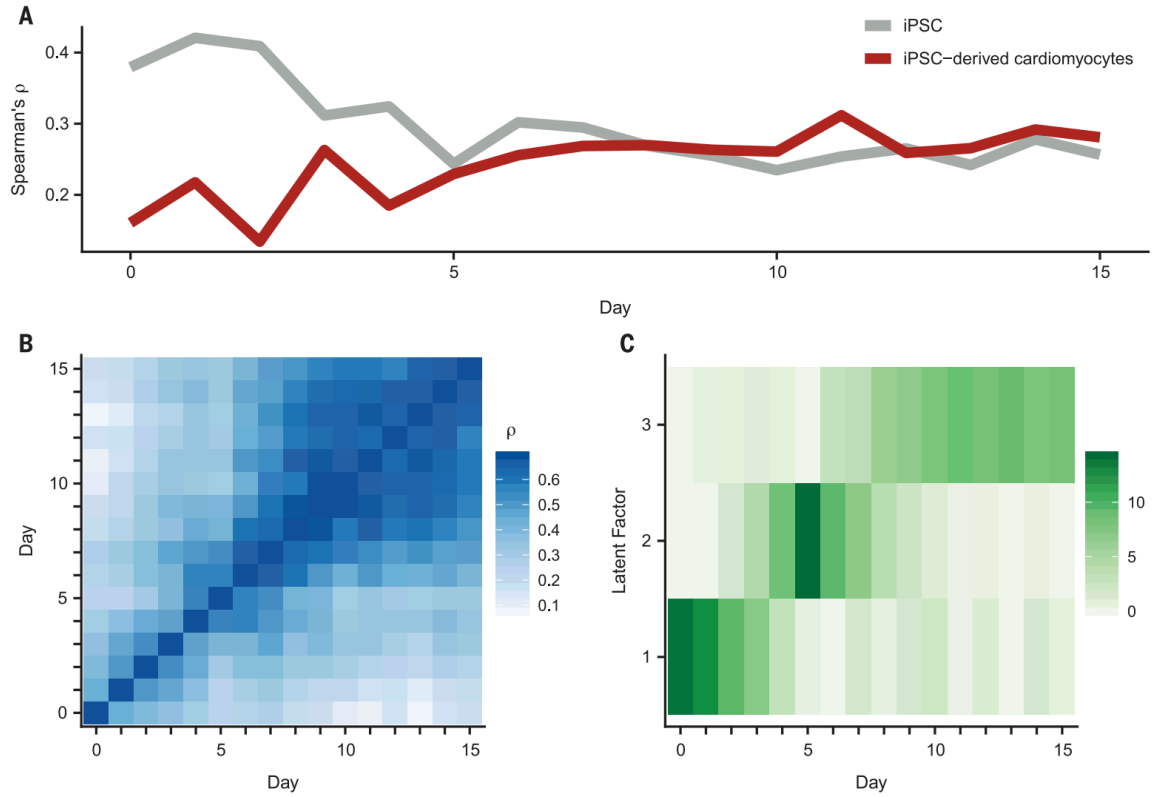


Fig. 2-2. eQTL patterns during cardiomyocyte differentiation. We limit this analysis to genes with at least one significant eQTL (WASP combined haplotype test; eFDR ≤ 0.05) across time points. If a gene has more than one significant eQTL, we select a single variant for that gene with the smallest geometric 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.

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 (Ernst et al. 2017, Roadmap Epigenomics Consortium 2015). 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) (Liberzon et al. 2015) and also show significant enrichment for genes related to dilated cardiomyopathy (P = 0.001, Fisher's exact) (table S2-7) (*Materials and Methods*, Burke et al. 2016). 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) (Hong et al. 2014, Arking et al. 2014). 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 (Rook et al. 2012). 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 (Ernst et al. 2017, Roadmap Epigenomics Consortium 2015). 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 (GTEx Consortium 2017).

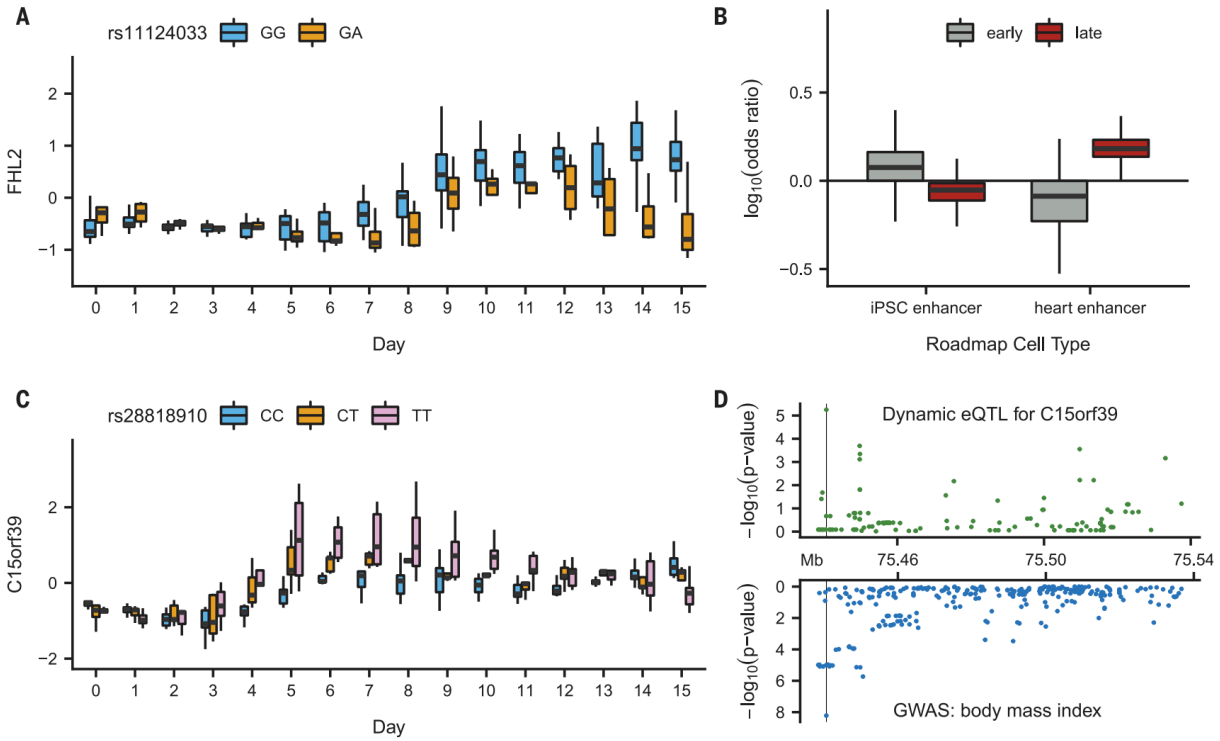


Fig. 2-3. Dynamic eQTLs detect genetic regulatory changes caused by cardiomyocyte differentiation. (A) Linear interaction association between genotype (color) of rs11124033 and time point (x axis) on residual gene expression (cell line effects regressed on expression) of *FHL2* (y axis). (B) Enrichment 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 (Zhou et al. 2016), 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) (Churchhouse et al. 2017) and weakly associated with red blood cell count ($P < 1.48 \times 10{-}6$, reported) (Astle et al. 2016). This dynamic eQTL and both traits show similar patterns of association across the region (fig. S2-27). The rs28818910 locus is associated with interindividual 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 (Banovich et al. 2018). 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 (Banovich et al. 2018). 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 cardiacspecific marker, cardiac Troponin T (564767, BD Biosciences) (Table S2-2). Cells were profiled on the BD LSRFortessa 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 (Degner et al. 2012).

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 (Banovich et al. 2018) and replication of day 15 cell lines within previously generated iPSC-derived cardiomyocyte cell lines (Banovich et al. 2018). 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 (Lazaro-Gredilla et al. 2012) 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 (Hensman et al. 2015).

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) (van de Geijn et al. 2015) 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) (Gamazon et al. 2013) 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` (Pedregosa et al. 2011). 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 $\left|\beta_{vg}(t=0)\right| \geq \left|\beta_{vg}(t=15)\right|$
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.

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 (Ernst et al. 2017). 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 (Roadmap Epigenomics Consortium 2015):

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 (Roadmap Epigenomics Consortium 2015):

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: Dynamic genetic regulation of gene expression during cardiomyocyte differentiation using single-cell RNA-seq

Abstract

During a dynamic process such as cellular differentiation, overall cell type composition and gene regulation both experience significant changes over time, which can vary by individual. To distinguish between these dynamic effects, we collected single-cell RNA sequencing data over a differentiation time course from induced pluripotent stem cells to cardiomyocytes, capturing 7 unique time points in 19 human cell lines. We identified dynamic eQTLs whose effects vary significantly with differentiation time, and classified these dynamic effects as primarily representing changes in overall gene regulation, changes in gene regulation in one cell type, or changes in cell type composition over time. We found that using cells from only one particular trajectory, rather than the entire single-cell dataset, results in the discovery of many more dynamic eQTLs, including those enriched for cell-type-relevant phenotypes. Finally, we used cell type composition information from single-cell data to deconvolute matched bulk RNA-seq samples and identify additional dynamic eQTLs that were not found prior to cell type deconvolution using single-cell data.

Full Text

During cellular differentiation, a cell undergoes major changes in its gene expression profile as it transitions from one cell type to another (Zeitlinger et al. 2010; Yan et al. 2013). These changes are guided by the genetic regulation of gene expression in each cell, which may be specific to a particular developmental time point or environmental context (GTEx Consortium 2017; Knowles et al. 2017). Dynamic gene expression data can allow us to uncover genetic variants involved in the regulation of the genome during a developmental process, and also enable the identification of regulatory variants with transient effects that may not otherwise be found.

Previous studies have investigated the dynamics of gene regulation during a differentiation time course using bulk RNA-sequencing (Strober et al. 2019). However, some questions remain unanswered, and new questions arise using this approach. During cellular differentiation, changes in the gene expression profile of a cell culture sample may reflect true changes in the gene regulation of each cell, changes in the total cell type composition of the sample, or both (Trapnell et al. 2015). Bulk RNA-seq data provides an average expression value of all cells in a given sample, so cell-specific information that would help distinguish between cell type composition changes and true gene regulatory changes is lost. In contrast, single-cell RNA sequencing preserves cell-specific gene expression information, enabling us to profile the cell type composition of heterogeneous samples.

The resolution of single-cell gene expression data and the possibility of classifying cell types also allows us to study differentiation trajectories in greater detail. In a previous study, bulk RNA-seq data from a cardiomyocyte differentiation revealed that cell lines cluster into distinct groups that exhibit differences in the expression trajectories of groups of genes over time

(Strober et al. 2019). With only average expression measurements for each sample from bulk data, it is unclear what characterizes these cell line clusters or the distinct differentiation pathways they may represent. Single-cell RNA-sequencing may help address these questions by providing cell-specific information to enable the study of gene regulatory and cell type composition changes during the transition from a pluripotent to a terminal cell type.

To that end, we performed cardiomyocyte differentiation on induced pluripotent stem cells (iPSCs) from 19 human cell lines, and collected single-cell RNA-seq data using Drop-seq at 7 informative days throughout the process. Single-cell data collection was performed with a balanced study design in which each collection contained three individuals at three unique differentiation time points, to minimize technical batch effects associated with individual and differentiation day (Table S3-1). After processing, the resulting 133 samples contained an average of 1962 cells per sample, with mean 1001 genes detected per cell. Following cell filtering and normalization, a principal component analysis shows that differentiation day is a major contributor to variation in the data (Fig. S3-1, S3-2).

Unsupervised clustering of time course single-cell data shows that cells from the same differentiation day broadly cluster together, although these groups are not as distinct in later cardiomyocyte time points (days 7, 11, and 15) (Fig. 3-1A). Marker genes known to be expressed at various stages in cardiac differentiation, from iPSC to mesoderm to cardiomyocyte, show high expression at expected early, intermediate, and late stages in the single-cell data, respectively (Fig. 3-1B). Interestingly, there appear to be two mutually exclusive cell types in the later time points. Differential expression analysis shows that one of these cell type clusters has high expression of genes known to be involved in cardiomyocyte function, such as *TNNT2* and *MYL7* (Ahmad et al. 2008, Bizy et al 2013). The other dominant cell type cluster in later time

points shows high expression of genes such as *COL3A1* and *VIM*, which are expressed in the extracellular matrix of cardiac fibroblast (Ieda et al. 2009, Zhang et al. 2019). The differentiation outcome of each sample varies by individual cell line; in some cell lines, cells differentiate primarily into either the *TNNT2*-expressing or the *COL3A1*-expressing terminal cell type clusters (Fig. 3-1C, S3-3A).

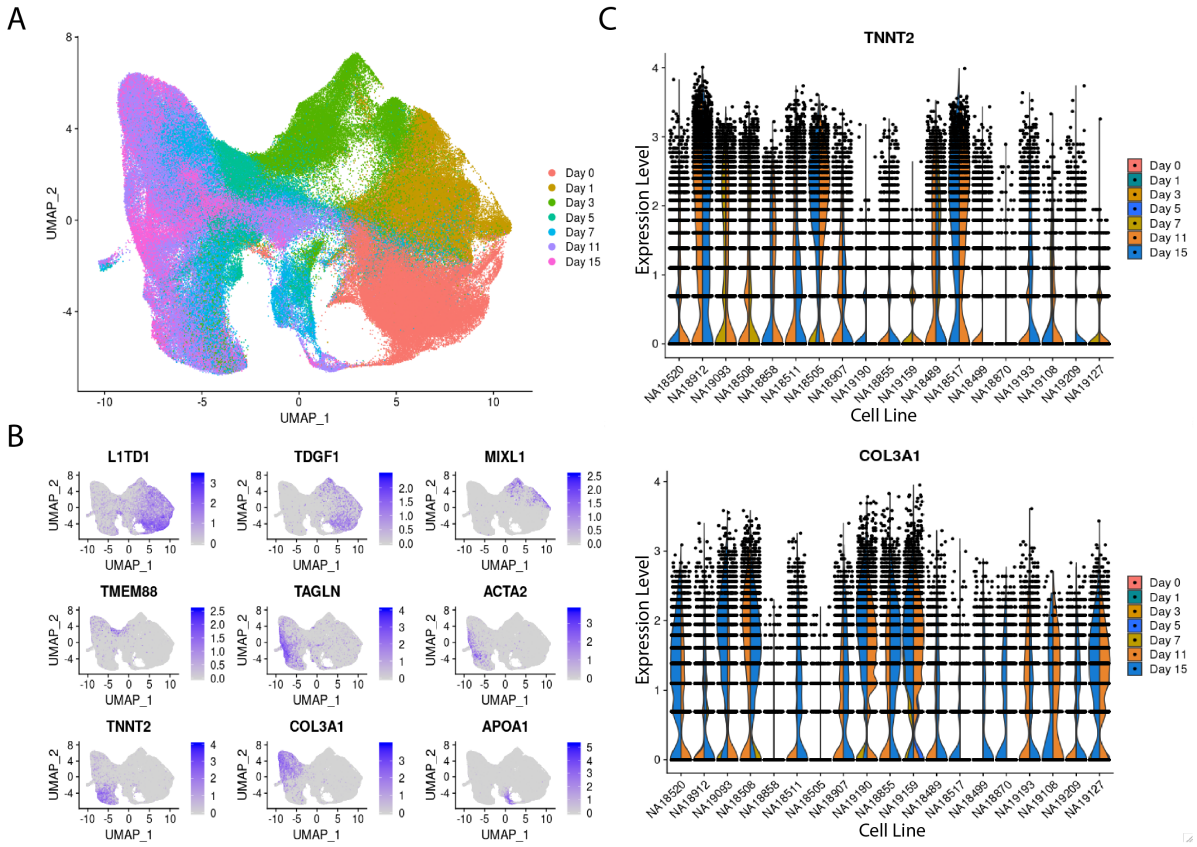


Fig. 3-1. Gene expression patterns in single cell data. (A) UMAP of full single cell dataset; cells are colored by differentiation day. (B) Normalized expression of marker genes on UMAP. Cells not expressing the gene are shown in gray. (C) Normalized expression level of *TNNT2* (top) and *COL3A1* (bottom) for each of 19 cell lines, colored by differentiation day. Cell lines are ordered in the same way in top and bottom panels.

Since bulk RNA-seq data was previously collected from the same cell lines during cardiomyocyte differentiation, we can directly compare these results by aggregating single-cell data into 'pseudobulk' RNA-seq data. Aggregated gene expression from earlier time points in

pseudobulk is most similar to data from earlier time points in bulk; and data from later time points in pseudobulk is most similar to data from later time points in bulk in the same cell lines (Fig. 3-2). The correlation of previously collected bulk gene expression data with single-cell pseudobulk expression data suggests that variation in gene expression profiles captures genetic factors, and does not merely reflect batch-specific technical effects.

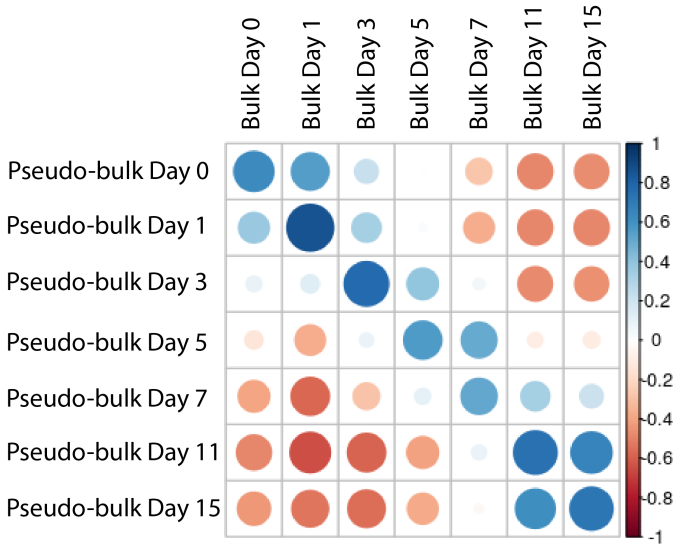


Fig. 3-2. Correlation of bulk and pseudobulk samples. For each differentiation day, we aggregated gene expression data across all 19 cell lines and computed Pearson correlation between single-cell pseudobulk data and bulk RNA-seq data collected from the same individuals in a cardiomyocyte differentiation time course (Strober et al. 2019).

Using the aggregated single-cell pseudobulk data, we used a Gaussian linear model to identify linear dynamic eQTLs whose effect varies significantly with time. This model identifies variant-gene pairs in which variant genotype and differentiation time significantly interact to affect gene expression, while controlling for the linear effects of both differentiation time and genotype. Using the full single-cell pseudobulk data, we identified 389 dynamic eQTL variants in 87 genes, at a local false sign rate below 0.05 using adaptive shrinkage (Stephens 2017). Genetic variants with these dynamic effects show significant enrichment for genes with roles in myogenesis, as well as genes related to dilated cardiomyopathy (P=0.018, Fisher's exact).

To further investigate the effect of cell type on these apparent gene regulatory effects, we identified linear dynamic eQTLs from the single-cell pseudobulk data solely for cells along the *TNNT2*-expressing trajectory or along the *COL3A1*-expressing trajectory, respectively (Fig. S3-3B). We identified 3440 dynamic eQTL variants in 398 genes in the *TNNT2*-trajectory cells alone, and 578 dynamic eQTL variants in 90 genes in the *COL3A1*-trajectory cells alone (lfsr < 0.05). Dynamic eQTLs identified using the *TNNT2*-trajectory cells show enrichment for genes related to myogenesis (P=1.2e-4), while eQTLs identified in the *COL3A1*-trajectory cells do not. It is notable that separating cells according to their specific differentiation trajectory results in a greater number of dynamic eQTLs compared to the full heterogeneous single-cell data. This is particularly true for the *TNNT2* trajectory, whose terminal cell type has an expression profile closer to the intended cardiomyocyte outcome of the differentiation protocol, and which produces many more dynamic eQTL variants upon separation from the full dataset.

Dynamic eQTLs identified thus far, particularly using the full single-cell dataset, may exhibit the observed changes to gene expression over time due to either gene regulatory differences or cell composition differences between individuals. One way to distinguish between these effects would be to remove the effect of cell type composition from the single-cell gene expression data, and examine only the residual expression values, which would be informative of gene regulatory differences over time.

Therefore, to separate the effect of gene regulation from cell composition, we regressed out a term for "cell type composition" derived from the single-cell data (as the proportion of cells in each unsupervised Seurat cluster per sample), and identified linear dynamic eQTLs using the residuals of this regression. We identified 255 dynamic eQTL variants in 52 genes using this cell-type-composition-residual data (lfsr < 0.05). Since the effect of cell type was removed, these

dynamic effects are likely to be primarily a result of true changes in gene regulation throughout the time course. Figure 3-3A shows an example of a dynamic gene regulatory effect, where an individual's genotype is significantly associated with differences in gene expression early in the time course, and the effect decreases over time.

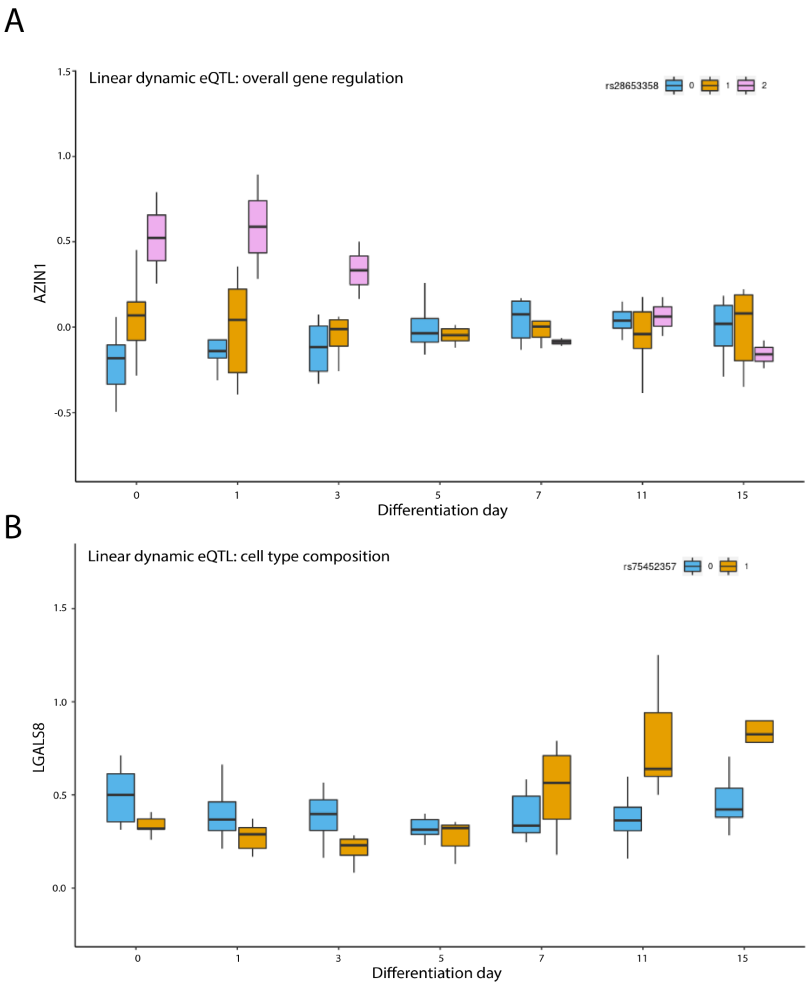


Fig. 3-3. Linear dynamic eQTL for gene regulation and cell type composition. (A) Linear interaction association between genotype (color) of rs28653358 and differentiation day (x axis) on residual gene expression of *AZIN1* (y axis). Cell type composition terms were regressed on expression prior to QTL mapping. (B) Linear interaction association between genotype (color) of rs75452357 and differentiation day (x axis) on gene expression of *LGALS8* (y axis).

By contrast, we can identify dynamic effects for which cell type composition is a major factor by investigating the variant-gene pairs that were found to be significant before the regression of the cell type composition term, but were not significant after regression.

Of these effects, a subset of dynamic eQTLs may be due to differences in cell type composition between individuals over time, and a subset may be due to true gene regulatory differences that only exist within one cell type, but are not captured in the entire dataset. We can identify dynamic genetic effects present in just one cell type by regressing out a "cell type" term for individual cell types (as the proportion of cells in that particular unsupervised Seurat cluster per sample); gene regulatory effects specific to that cell type will be removed from those identified using the residual gene expression values. Using this method, we were able to identify 77 genes with a dynamic eQTL effect specific to only one cell type, which were not identified as dynamic eQTLs affecting gene regulation across all cell types in the previous cell-type-composition regression model (lfsr < 0.05). Notably, 62 of these genes have a dynamic eQTL specific to a cluster with high expression of cardiac marker *TNNT2*.

Finally, loci identified by our dynamic eQTL model that were found after individual-cell-type regression but not found after cell-type-composition regression are likely to be dynamic genetic effects that primarily reflect changes in overall cell type composition over time. An example of a cell type composition dynamic eQTL is shown in figure 3-3B, where the genotype effect on gene expression increases as the differentiation progresses.

For the dynamic eQTLs identified thus far, we have investigated the interaction between genotype and chronological time, represented by the differentiation day in which each sample was collected. However, chronological time may not be the best measure of biological time for every sample. Different cells and/or individual cell lines may progress through the differentiation at different rates, which means the biological stage represented by "Day 3" for some individuals may not be the same stage as "Day 3" for others, for example. To account for this, we used PAGA to infer pseudotime values for each single cell, by reconstructing branching gene

expression changes across the dataset (Wolf et al. 2019) (Fig. 3-4). Using these pseudotime values instead of chronological time, we found 572 linear dynamic eQTL variants for 77 unique eGenes (lfsr < 0.05). These pseudotime linear dynamic eQTLs may represent dynamic genetic effects arise more gradually over time, as pseudotime values provide a continuous gradient compared to discrete chronological time points. Only a minority of these pseudotime dynamic eQTLs (11.5%) were also identified in the previous model using chronological time, suggesting that building a continuous pseudotime trajectory enhances our ability to detect dynamic effects over the time course.

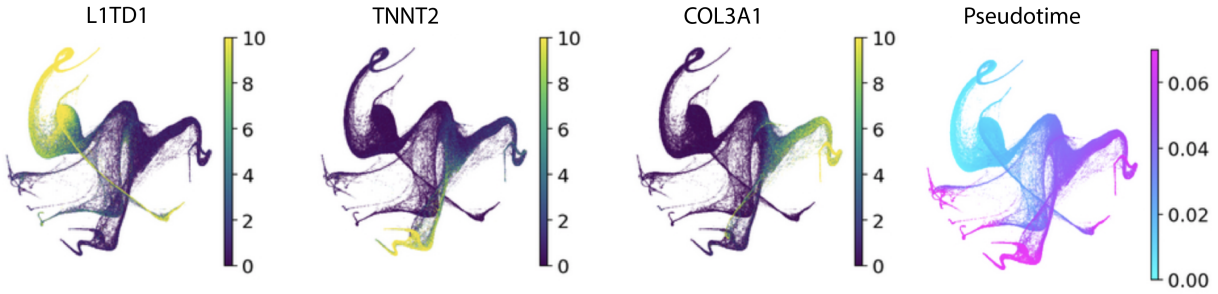


Fig. 3-4. Pseudotime trajectories in single cell data. Partition-based graph abstraction (PAGA) showing the expression levels of one early time course marker gene (*L1TD1*) and two terminal cell marker genes (*TNNT2* and *COL3A1*) along an inferred pseudotime trajectory, whose values are shown on the right.

Since pseudotime values can infer intermediate time points with more resolution than chronological time, we used these values to identify nonlinear dynamic eQTLs, or eQTLs whose effect is present in the middle of the differentiation, but not at the beginning or the end. We identified 2849 nonlinear dynamic eQTL variants for 355 eGenes using inferred pseudotime values (lfsr < 0.05). Of these 355 genes with a nonlinear dynamic effect, only 17 were also identified as having a linear dynamic eQTL using chronological time points. Our time course study design is particularly useful for detecting these nonlinear genetic effects, as they may be transient and would not be found by studying only the initial or terminal cell types of a dynamic process such as differentiation.

Using the information on cell type composition from single-cell data, we can return to previously published bulk RNA-seq data of the same cell lines undergoing cardiomyocyte differentiation. Under the assumption that bulk RNA-seq data also contains a heterogeneous cell population, we used the cell type composition terms obtained from single-cell data to deconvolute the effect of cell type from the dynamic gene regulatory effects found in bulk. The inclusion of cell type composition terms to bulk data resulted in enhanced detection of linear dynamic eQTLs (compared to randomly sampled expression data with the same terms, $P < 2.2e-16$). These additional eQTLs represent dynamic effects that have been uncovered only due to data deconvolution using cell type composition terms from single-cell RNA-seq data.

We found that single-cell RNA-seq data collected during a differentiation time course can disentangle the effects of gene regulatory changes from cell composition changes over time. Our cardiac differentiation resulted in more than one terminal cell type expression profile, and specifically selecting for our trajectory of interest enabled us to identify many more dynamic effects that would otherwise have been obscured, even with the full single-cell dataset. The approach of determining cell type composition with single-cell data can be used to deconvolute cell type effects from gene regulatory effects in a matched bulk RNA-seq dataset. The dynamic genetic effects identified in this study may provide a resource for investigating mechanisms underlying developmental processes that include heterogeneous and dynamically changing cell types over time.

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 (Banovich et al. 2018). 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 (Banovich et al. 2018). 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 13 of differentiation.

In some cases, after performing cardiac differentiation, one might choose to perform a post hoc purification process to remove any non-cardiac cell types present at the terminal time point (Tohyama et al. 2013). However, for the purposes of a time course experiment where multiple intermediate time points are assayed, a purification protocol undertaken only at the end of the differentiation would not prove useful; therefore, no cell type purification was performed.

Sample Collection and Processing. We performed cardiomyocyte differentiations in three total batches of six to seven cell lines at a time. For each batch, cardiomyocyte differentiations were performed with three staggered starting days, such that samples could be collected from each cell line in three differentiation stages at any given time. For all 19 cell lines, samples were collected on differentiation days 0 (iPSC, before treatment with CHIR99021), 1, 3, 5, 7, 11, and 15. Drop-seq collection was performed a total of three collection days for each batch of six to seven cell lines. In the first collection day, samples from all cell lines in the batch were collected for differentiation days 1, 3, and 7. In the second collection day, samples from all cell lines in the batch were collected for differentiation days 5, 7, and 11. In the third collection day, samples from all cell lines in the batch were collected for differentiation days 0 (iPSC), 11, and 15. Through this process, single-cell gene expression data was collected for all cell lines in seven unique time points, with two time points (differentiation days 7 and 11) having two replicates.

This staggered differentiation and collection study design was performed to minimize the technical effect of sample collection as a potential confounding variable associated with cell line or differentiation day.

To harvest the samples at the start of each collection day, cells in at least two wells of a 6-well culture dish were released from the dish using Accutase (BD Biosciences, #561527). Samples were washed three times and resuspended in 1X PBS, 0.01% BSA. Cells were then passed through a 40 um filter to encourage the formation of a single cell suspension. The concentration of each single cell suspension was quantified manually using an NI hemocytometer (InCyto, DHC-N01-2).

Using a 125 um Drop-seq microfluidic device, single cells were captured in droplets along with a DNA barcoded bead (ChemGenes, Macosko-2011-10(V+)), following the standard Drop-seq protocol (Macosko et. al 2015). The DNA barcoded beads include a cell-specific barcode so the cell identity of each RNA molecule can be recovered. After Drop-seq collection, the RNA molecules were reverse transcribed, and cDNA amplification was performed according to the Drop-seq protocol. cDNA concentration and library size were measured using the Qubit 3 fluorometer (Thermo Fisher) and BioAnalyzer High Sensitivity Chip (Agilent, #5067-4626).

Library preparation was performed using the Illumina Nextera XT DNA Library Preparation Kit (Illumina, FC-131-1096). Libraries in each batch were multiplexed together so that every sequencing lane contained three samples, one from each of the three collection days. Each of those samples was itself a multiplexed collection of three individual cell lines at three distinct differentiation time points, which were mixed upon Drop-seq collection. Samples went through paired-end sequencing using the Illumina NextSeq 500. 20 bp were sequenced for Read

1, and 60 bp for Read 2 using Custom Read 1 primer, GCCTGTCCGCGGAAGCAGTGGTATCAACGCAGAGTAC, according to manufacturer's instructions (Macosko et al. 2015). The same multiplexed library pool was sequenced twice with the goal of achieving at least 20 million reads per sample

We recorded 20 technical and biological covariates and measured their contribution to variation in our data (Fig. S3-2).

Genotype data. We used previously collected and imputed genotype data for the 19 Yoruba individuals from the HapMap and 1000 Genomes Project (Degner et al. 2012).

RNA-seq quantification. For each sequencing run, we obtained paired-end reads, with one pair representing the cell-specific barcode and unique molecular identifier (UMI), and the second pair representing a 60 bp mRNA fragment. We used dropseqRunner (available at github.com/aselewa/dropseqRunner) which takes a fastq file with paired-end reads as input and produces an expression matrix corresponding to the UMI of each gene in each cell. All RNA-seq samples were aligned to the human genome (GRCh38) using STAR-solo (Dobin et al. 2013). We used *featureCounts* (Liao et al. 2014) to assign each aligned read to a genomic feature, and *umi_tools* (Smith et al. 2017) to create a count matrix representing the frequency of each feature in our dataset.

60,668 genes were used for downstream analysis. Single-cell data were filtered to remove cells with fewer than 200 genes per cell, percent mitochondrial reads greater than 30%, or doublet probability greater than 30% according to the single-cell demultiplexing software demuxlet (Kang et al. 2018). The data was normalized using the *Seurat* R package's sctransform

function, following the standard pipeline recommended by Seurat (Stuart et al. 2019). All downstream analysis used gene expression data following sctransform normalization and scaling.

Dimensionality reduction. To perform principal components analysis (PCA), Seurat first identifies a subset of highly variable genes by calculating gene dispersion compared to mean expression values for each gene. These highly variable genes are used to calculate principal components for the single-cell data. The top 30 global expression PCs were then used to perform clustering and visualization on a Uniform Manifold Approximation and Projection (UMAP). Seurat's 'FindClusters' function was used to perform unsupervised clustering on single-cell data using a resolution of 0.65.

Cell type trajectory categories. Seurat clusters were categorized as belonging primarily to *TNNT2*-trajectory, primarily to *COL3A1*-trajectory, to both, or having no particular pattern. Clusters that showed expression of *TNNT2* without expression of *COL3A1* were placed in the *TNNT2* category, as well as earlier time point clusters that had high contact with these clusters according to UMAP visualization. Conversely, clusters that showed expression of *COL3A1* without expression of *TNNT2* were placed in the *COL3A1* category, along with earlier clusters leading into those clusters according to UMAP visualization. Seurat clusters representing the earliest time points did not show a clear visual distinction between trajectories, and were placed in both the *TNNT2*-trajectory and *COL3A1*-trajectory categories. Seurat clusters that did not show a clear pattern or contained cells scattered along both lineages were also placed in both trajectory categories. In all, the *TNNT2*-trajectory and *COL3A1*-trajectory groups shared cells from 14 Seurat clusters; the *TNNT2*-trajectory group uniquely contained cells from 4 Seurat clusters; and the *COL3A1*-trajectory group uniquely contained cells from 5 Seurat clusters (Fig. S3-3B).

Correlation between bulk and pseudobulk data. Single cell data was aggregated into pseudobulk by taking the sum of normalized gene expression values across all cells of a given sample for a given gene. For downstream analysis that uses single-cell data in pseudobulk, including correlation with bulk RNA-seq data and pseudobulk dynamic eQTL calling, we filtered samples to exclude those with fewer than 700 cells per sample. We used normalized expression (gene by sample) matrices from both bulk RNA-seq and pseudobulk data, using genes that were detected in both datasets. We calculated the Pearson correlation of the normalized gene expression matrix from bulk RNA-seq data (collected in *Chapter II*) with the normalized gene expression matrix from the pseudobulk RNA-seq data (calculated from single-cell data, collected in *Chapter III*).

Cell line collapsed PCA. We used a "cell line collapsed PCA" approach to identify broad, cell line specific patterns across the entire time course. To identify cell line collapsed PCs, we rearranged the gene expression matrix from the standard RNA-seq quantification such that each row represented expression from one cell line and each column represented a gene at a single time point. 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. We controlled for the first five cell line collapsed PCs when detecting dynamic eQTLs.

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).

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:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \cdots + \beta_7 PC_c^5 + \beta_8 G_c t, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and time (β_8). To determine significant effect size difference, we used the R package 'ashR' which performs adaptive shrinkage, a flexible empirical Bayes approach using estimates of the coefficient of interest and its standard error from the linear model (Stephens 2017). From this adaptive shrinkage approach, we use the measure lfsr, or local false sign rate, to determine significance of each variant gene pair.

We tested the same set of variant-gene pairs that was tested in the dynamic eQTL calling analysis in *Chapter II* using bulk RNA-seq data (with the exception of genes from that dataset that were not detected in the single-cell data). This was done to reduce bias when comparing eQTLs from these two analyses.

Classifying linear dynamic eQTLs in terms of gene regulation or cell type composition. To identify eQTLs that are primarily associated with overall gene regulation, we regressed out terms that represent cell type composition per sample. There were a total of 23 terms corresponding to the 23 clusters identified using unsupervised Seurat clustering (resolution 0.65). Each of these 23 terms contained values for all pseudobulk samples that represented the proportion of cells in that

cluster per sample using single cell expression data. We regressed out all 23 cell type terms from the standard gene expression matrix per sample. The residuals from this linear model were then used in place of pseudobulk gene expression values to identify linear dynamic eQTLs using the general linear dynamic eQTL model above. We called dynamic eQTLs identified using the residuals from the overall cell-type-composition model as eQTLs related to overall gene regulation, since the effect of cell type composition has been removed through this process.

To distinguish between dynamic eQTLs that primarily affect cell type composition over time, and dynamic eQTLs primarily associated with gene regulation but only in one cell type, we performed multiple linear regressions, each using only one of the 23 cluster proportion terms referenced above. The residuals from these linear models were used in place of pseudobulk gene expression values to identify linear dynamic eQTLs using the general linear dynamic eQTL model above. We called dynamic eQTLs that were identified using the residuals from each single-cluster term, and were also identified using the standard linear dynamic eQTL model with pseudobulk expression values, as eQTLs related to general cell type composition, since removing the effect of single cell types did not remove the eQTL effect. By contrast, we called dynamic eQTLs that were identified using the standard expression value dynamic eQTL model, but were not identified using the single-cluster-residual model, as eQTLs related to gene regulation in that particular cluster or cell type, since removing the effect of that cluster or cell type abolished the eQTL effect.

For all linear dynamic eQTL models in this section, we used R `lm` to quantify the significance of the interaction between genotype and time. We used the R package 'ashR' which performs adaptive shrinkage using estimates of the coefficient of interest and its standard error from the linear model. From this adaptive shrinkage approach, we use the measure lfsr, or local

false sign rate, to determine significance (Stephens 2017). We tested the same set of variant-gene pairs that were tested in the standard linear dynamic eQTL model described in the section above.

Bulk dataset deconvolution using single cell data. We used information on cell type composition from the single cell data to deconvolute the dataset previously collected on the same cell lines using the same differentiation protocol with bulk RNA-seq (Strober et al. 2019). Since the bulk dataset contained samples collected at more time points throughout the differentiation, we limited our analysis to only those bulk samples that had a matching sample in the single cell data. We detected linear dynamic eQTLs using the previously published bulk RNA-seq data first with the same Gaussian linear model used previously (i.e. without any cell type deconvolution terms). We then detected linear dynamic eQTLs using the bulk RNA-seq data with a Gaussian linear model with added cell type composition terms. These were a total of 23 terms corresponding to the 23 unsupervised Seurat clusters identified in the matched single-cell RNA-seq data. Each of these 23 terms contained values for all matched bulk samples that represented the proportion of cells in that cluster per matched single-cell sample, calculated using single-cell expression data. This bulk deconvolution model was run as follows:

We built a separate linear model for each tested variant-gene pair. 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, $PC^{K} \in R^{C}$ denotes the Kth cell line collapsed PC vector, and Z_I denotes the proportion of cells in the Ith cluster. We modeled the expression levels as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \beta_4 PC_c^1 t + \cdots + \beta_{11} PC_c^5 + \beta_{12} PC_c^5 t + \beta_{13} Z_1 + \beta_{14} Z_2 + \cdots + \beta_{35} Z_{23} + \beta_{36} G_c t, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and time. We then used the R package 'ashR' which performs adaptive shrinkage, and calculated local false sign rate to determine significance (Stephens 2017). We tested the same set of variant-gene pairs that was tested in the dynamic eQTL analysis in *Chapter II* using bulk RNA-seq data.

Pseudotime trajectory analysis using PAGA. To perform pseudotime trajectory analysis, we imported the full single-cell Seurat object (post-normalization) to scanpy and followed the standard PAGA pipeline (Wolf et al. 2018, Wolf et al. 2019) . We used Leiden clustering at 0.03 resolution to cluster the single-cell data in an unsupervised manner. Leiden clusters were then used to initialize PAGA and compute pseudotime trajectories.

From PAGA, pseudotime values were obtained for every cell in the single-cell data. To use this information in dynamic eQTL models which use pseudobulk data, we grouped single cells by sample (where a sample refers to a given cell line at a given differentiation day) and obtained the median pseudotime value of all cells in that sample. These median pseudotime expression values per sample were then used in place of chronological differentiation day for the purposes of pseudotime dynamic eQTL mapping.

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:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 t^2 + \beta_4 PC_c^1 + \cdots + \beta_8 PC_c^5 + \beta_9 G_c t + \beta_{10} G_c t^2, \sigma)$$

We used R `lm` to quantify the significance of the interaction between genotype and the second order time variable (β_{10}). We then used the 'ashR' to perform Bayesian adaptive shrinkage, using estimates of β_{10} and its standard error from the linear model (Stephens 2017). From this adaptive shrinkage approach, we use the measure lfsr, or local false sign rate, to determine significance of each variant gene pair.

Supplementary Figures and Tables

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

Chapter IV: Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation

Note:

The following section (*Chapter IV*) is a summary of a project to which I contributed, titled "Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation" (Selewa et al. 2020). This paper was published in *Scientific Reports* on January 30, 2020. My contribution to the project is detailed below.

Authors:

A. Selewa, R. Dohn, H. Eckart, S. Lozano, B. Xie, E. Gauchat, R. Elorbany, K. Rhodes, J. Burnett, Y. Gilad, S. Pott, and A. Basu

Introduction

The development of single-cell sequencing technologies has greatly enhanced our ability to study transcriptomes at high resolution. To ensure quality transcriptomic data, single-cell technologies such as Drop-seq require suspensions of intact, mostly viable single cells (Macosko et al. 2015). However, there are many cases in which intact single cell suspensions cannot be obtained -- for example, for cell types with unusual morphology that can't reliably pass through a single cell filter, or cells or tissues where the cell membrane has been breached during handling. In these cases, a useful alternative may be to obtain gene expression profiles from single nuclei, which do not require intact viable single cells. DroNc-seq is a microfluidic approach similar to Drop-seq that enables the collection of single-nucleus, rather than single-cell, gene expression data (Habib et al. 2017). To establish a greater understanding of the strengths and limitations of these two approaches, we performed a systematic comparison of Drop-seq and DroNc-seq during a differentiation time course from human iPSCs to cardiomyocytes (Selewa et al. 2020).

For this project, I maintained and differentiated two human cell lines from induced pluripotent stem cells to iPSC-derived cardiomyocytes. Samples from these cell lines were collected by Drop-seq for single-cell information and DroNc-seq for single-nucleus information on days 0 (iPSC), 1, 3, 7, and 15 of the differentiation protocol.

Summary of results

The study found that single-nucleus expression data from DroNc-seq yields broadly similar results to single-cell data from Drop-seq on matched samples throughout the time course. DroNc-seq captured a significantly higher proportion of intronic reads compared with Drop-seq,

likely because these reads come from unprocessed RNA molecules which are enriched in the nucleus (Selewa et al. 2020). Cell type-specific genes were identified using unsupervised Seurat clustering, by the top marker genes for each cluster. Clusters identified by both Drop-seq and DroNc-seq captured the anticipated gene expression trajectory throughout the differentiation from iPSCs to cardiomyocytes -- broadly expressing first pluripotency markers, followed by mesoderm and cardiac progenitor markers, followed by cardiomyocyte markers (Fig 4-1C). Interestingly, in addition to the cardiomyocyte-lineage cells, unsupervised clustering also detected cells in two smaller alternative-lineage clusters in later time points of the differentiation (Fig 4-1A,B) . These 'alternative lineage' cells may represent cells at intermediate stages of cardiac differentiation, cells that failed to differentiate, or cells that differentiated towards an alternative trajectory.

The study found that iPSC pseudo-bulk samples aggregated from single-cell and single-nucleus data were most highly correlated with bulk gene expression data from iPSCs collected in a previous study. By contrast, pseudo-bulk samples representing iPSC-derived cardiomyocytes from single-cell and single-nucleus data were most highly correlated with bulk gene expression data from iPSC-derived cardiomyocytes and primary heart tissue collected in a previous study (Fig 4-1E). Drop-seq and DroNc-seq expression data were used to infer a pseudo-time differentiation trajectory, which places iPSCs at the beginning of the trajectory, followed by cardiac progenitors, followed by cardiomyocytes (Fig. 4-1H,I). Finally, DroNc-seq was applied to frozen heart tissue to demonstrate the use of single-nucleus gene expression data to characterize cell type composition in frozen tissue as well as fresh cell culture samples. In all, this systematic comparison of Drop-seq and DroNc-seq demonstrates the ability of both

technologies to extract relevant gene expression information in a dynamic and heterogeneous context such as cardiomyocyte differentiation.

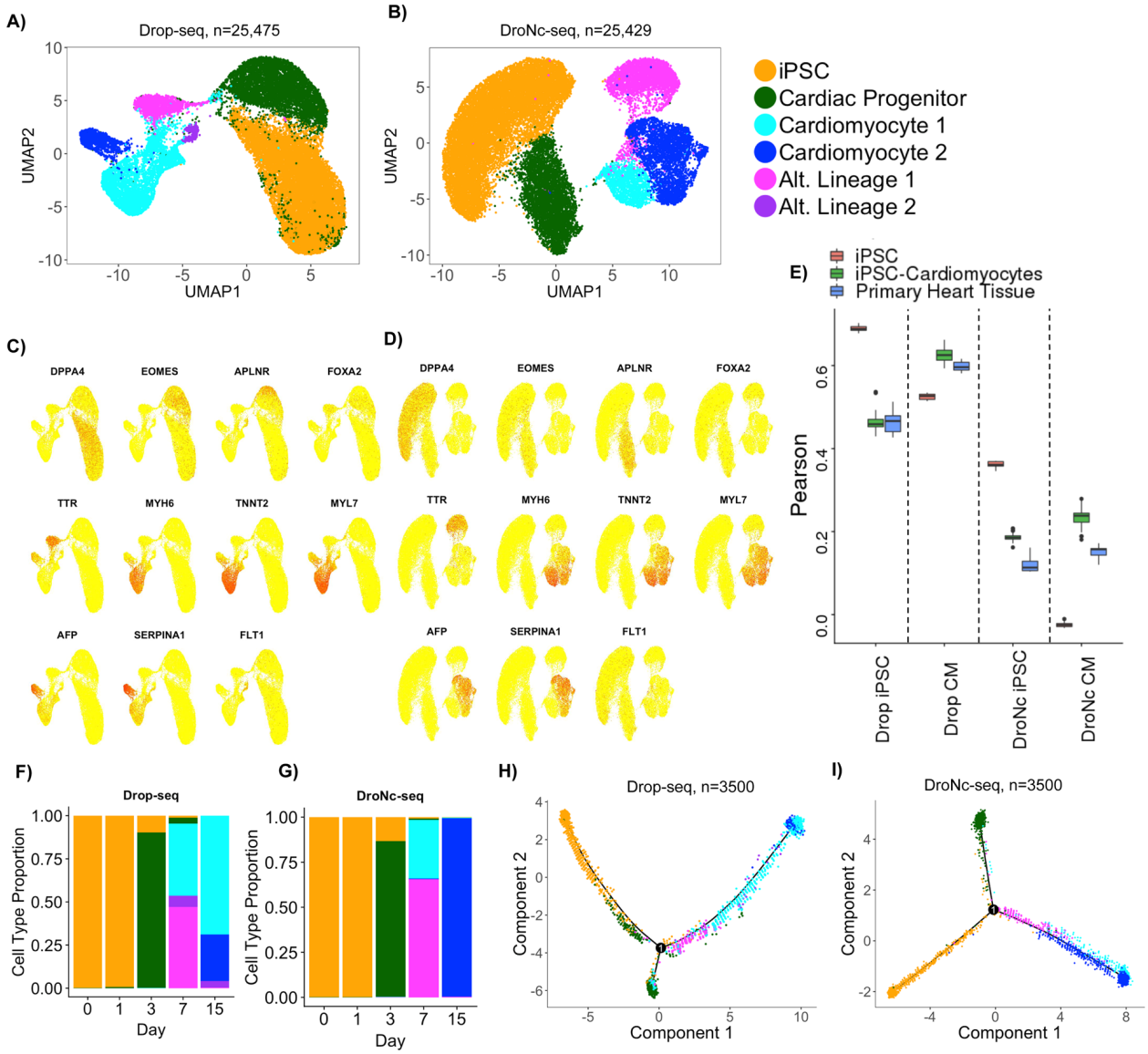


Fig. 4-1. Cell type and single-cell trajectory analysis from Drop-seq and DroNc-seq data. (A,B) Clustering results visualized with UMAP and colored by inferred cell type for Drop-seq and DroNc-seq. (C,D) Expression of marker genes overlaid on UMAP plots from A and B for Drop-seq and DroNc-seq. (E) Pearson correlation of DroNc-seq and Drop-seq pseudo-bulk against bulk RNA-seq from iPSCs (n = 18), iPSC-Cardiomyocytes (n = 51), and primary heart tissue (n = 22) (Pavlovic et al. 2018). (F,G) Distribution of cell types per time-point in Drop-seq and DroNc-seq, respectively. (H,I) Inferred trajectories using Monocle with color representing inferred cell types. A total of 3500 cells were used for the trajectory corresponding to 700 per time-point.

Discussion

This project was conducted before the experiments in *Chapter III*, in which I collected single-cell RNA-seq data during a cardiomyocyte differentiation for the purpose of studying the dynamics of gene regulation. The results of this study demonstrated the utility of Drop-seq in the collection of single cells in various stages of the differentiation protocol, from iPSCs at day 0 to cardiomyocytes at day 15.

This study also provided a clearer picture of which time points throughout the differentiation would be most informative in the collection of single-cell RNA-seq data for the time course performed in *Chapter III*. In addition to the time points included in this study (days 0, 1, 3, 7, and 15), collection at intermediate differentiation days 5 and 11 were added in *Chapter III* to increase the overlap of cell types between adjacent time points, as there did not seem to be significant overlap between cell types otherwise (Fig 4-1F, G). Detection of similar cell types across time-points would allow for the reconstruction of a more continuous differentiation trajectory, such as in a pseudotime analysis, to characterize the dynamic temporal relationship between cell types.

Interestingly, this study also found several distinct differentiation trajectories, as evidenced by the unsupervised clustering of two "alternative lineage" cells in addition to the cardiomyocyte-lineage cells. Given this result, this study demonstrated the ability of Drop-seq to identify single cells that might progress down distinct pathways using the same differentiation protocol, and provided more confidence in the possibility of analyzing cell type composition changes over time, as we proceeded to do in *Chapter III*.

Chapter V: Discussion

These combined projects demonstrate our ability to identify dynamic genetic effects on gene expression using a differentiation time course. Using induced pluripotent stem cells and their derived terminal cell types, we can identify genetic effects related to the interaction between an individual's genotype and the differentiation progress or cell type of a sample. We located dynamic eQTLs with relevance in heart-related phenotypes which had not previously been characterized. We also identified dynamic eQTLs that have a fleeting effect during intermediate time points, which would likely not have been found without the use of a differentiation time course. We used single-cell RNA sequencing to distinguish between the dynamic effects of gene regulation from the effects of cell type composition changes over time. We found that specifically selecting cells only within the differentiation trajectory of interest enables the identification of many more genetic effects, which may have been obscured by cell type heterogeneity in a full dataset. Finally, we were able to use cell type composition information captured from single-cell RNA-seq data to deconvolute matched bulk RNA-seq samples collected during cardiomyocyte differentiation.

The fact that selecting specifically for *TNNT2*-trajectory cells or *COL3A1*-trajectory cells results in the identification of many more dynamic eQTLs compared to the full single-cell dataset (*Chapter III*) suggests that heterogeneity in the full dataset may impede the identification of gene regulatory effects. We expected this to be the case using bulk RNA-seq data, which provides a single average expression measurement of all cells in a given sample, but found this to be true using all cells in a heterogeneous population using single-cell RNA-seq data as well. Collecting single-cell RNA-seq data, therefore, does not necessarily resolve complexities related to cell type heterogeneity in itself, without additional downstream cell type or trajectory disaggregation.

Although we could not identify distinct terminal cell types using bulk RNA-seq data, one might be tempted to compare the differentiation trajectories found in single-cell time course data, with the distinct cell line clusters observed in the bulk time course data. With bulk RNA-seq data collected during the differentiation time course in *Chapter II*, we used an unsupervised model, split GPM, that grouped cell lines into two clusters that exhibit broad differences in the expression trajectory of groups of genes over time. Our intention was to use single-cell RNA sequencing to provide insight into what might characterize these cell line clusters, with a possible explanation being that they represent distinct differentiation trajectories. Single-cell data collected during the differentiation time course (*Chapter III*) did indeed show that cell lines proceed down at least two distinct differentiation pathways, ending with terminal cells that highly express either *TNNT2* or *COL3A1*, respectively. However, the cell line clusters found using the split GPM model with bulk RNA-seq data do not closely correspond to groups of cell lines that end with either primarily *TNNT2*-expressing cells or *COL3A1*-expressing cells according to the single-cell data (data not shown).

There could be a number of explanations for this discrepancy. It may be that cell lines do not consistently follow either one differentiation pathway or the other, but that the outcome of differentiation varies from differentiation batch to batch due to environmental factors. It is likely that environmental factors do influence a cell line's differentiation trajectory, but there is also evidence that genetic factors play a role (Cuomo et al. 2020). The relatively high correlation between gene expression from bulk RNA-seq data and matched single-cell pseudobulk samples, which were differentiated and collected in two separate studies, suggests that genetic factors related to cell line may provide some consistency to the differentiation process.

Another explanation for the discrepancy between cell line clusters identified in the bulk RNA-seq and single-cell RNA-seq projects may be the resolution of differences that these groupings are identifying. Instead of distinguishing between cell lines progressing through the *TNNT2*-expressing trajectory or the *COL3A1*-expressing trajectory, cell line clusters identified using bulk RNA-seq could be picking up on one group of cell lines whose samples primarily end up in any cardiac lineage (including both *TNNT2*- and *COL3A1*-trajectory cells), and another group of cell lines where a relatively larger proportion of cells either diverge into a non-mesoderm pathway or fail to differentiate entirely. In single-cell data in *Chapter III*, there appears to be a third, smaller differentiation trajectory which ends in cells expressing genes such as APOA1 and AFP, according to unsupervised clustering (Fig. 4B). The genes characterizing this smaller trajectory are largely unrelated to cardiac differentiation, and may be characteristic of cell types in the endodermal lineage (Elshourbagy et al. 1985, Lazarevich et al 2000). Without cell type information from bulk RNA-seq data, we may conjecture that bulk samples similarly contained off-target results of differentiation such as these cell types, which may have contributed to their cell line cluster categorization according to split GPM.

All together, the results from these projects demonstrate the benefit of using a balanced time course study design to investigate dynamic gene regulatory differences between individuals, which can provide new insight into the potential role of genomic variants related to human development and disease.

However, these results invite further study. The identification of expression QTLs with a dynamic effect during cardiomyocyte differentiation provides many candidates for casual genomic loci potentially involved in the development of downstream phenotypes or disease. For some of these loci, we have gathered more evidence of their possible function and relevance by investigating their location in the genome as compared to previously annotated cis-regulatory elements. For example, the discovery that a dynamic eQTL is located in a promoter or enhancer region suggests that this locus may be involved in the function of those cis-regulatory elements on their associated protein-coding gene. We have also investigated the potential phenotype and disease relevance of dynamic expression loci identified during cardiomyocyte differentiation by investigating their overlap with disease-associated loci, such as those previously identified by genome-wide association studies. Nonetheless, further follow-up studies should be performed to validate the function of these genomic loci and their potential relevance to downstream phenotypes. One way to perform this functional validation may be to disrupt the function of candidate genomic loci using CRISPR-Cas9, and test whether this disruption causes a significant change in gene expression or other downstream molecular phenotypes (Gasperini et al. 2019).

We can also use orthogonal molecular phenotypes besides gene expression to investigate the potential mechanisms of action of our dynamic expression QTLs. In conjunction with the discovery of dynamic expression QTLs, we can investigate whether these loci are also correlated with changes in other molecular phenotypes such as chromatin accessibility, DNA methylation,

or RNA splicing, as previous studies have done (Degner et al. 2012; Li et al. 2016). An advantage of using induced pluripotent stem cells as a model system is the ability to perform multiple orthogonal assays using the exact same individuals and genotypes. During the cardiomyocyte differentiations performed for both the bulk RNA-seq (*Chapter II*) and single-cell RNA-seq (*Chapter III*) projects, leftover samples were collected and stored, enabling the possibility of using these same samples as input for ATAC-seq or bisulfite sequencing, for example. Using matched data for the same time points and in the same individuals, we might be able to see patterns between these molecular phenotypes over time that could suggest a functional relationship. For example, if changes in chromatin accessibility throughout the differentiation time course are often closely followed in time by changes in gene expression at the associated coding region, this provides evidence for a potential mechanistic effect of this dynamic QTL, in which changes in chromatin accessibility ultimately influence subsequent gene expression changes. The addition of multiple orthogonal molecular phenotypes to this study could greatly enhance our understanding of the effect and possible mechanism of dynamic gene regulatory effects during a differentiation time course.

In addition to these complementary approaches, new questions have emerged as a result of these studies. Of particular interest is the interpretation of the dynamic gene regulatory effects found at intermediate stages of the differentiation time course, but not at the beginning (in iPSCs) or at the end (in terminal cell types). Using bulk RNA-seq data, 25 genes with these intermediate dynamic eQTLs were identified, including some loci that overlap phenotype-relevant GWAS hits, such as the nonlinear dynamic eQTL for the gene C15orf39, which has been associated with body mass index (Churchhouse et al. 2017). The interpretation of these intermediate dynamic eQTLs is not immediately obvious. It may be the case that these effects are

the result of a cell type sub-population whose population size peaks at intermediate time points; thus, this is when we are most likely to find an effect, unobscured by other cell types in the sample. If so, we may be able to identify this effect using single-cell RNA-seq by specifically selecting the trajectory of these cells in which we see an intermediate dynamic eQTL effect, and investigating whether the effect is replicated across the entire time course when looking only at this cell population. On the other hand, these intermediate dynamic eQTLs may truly be fleeting effects that emerge momentarily throughout the differentiation, and then subside as the differentiation progresses. In this case, these intermediate dynamic eQTLs may not necessarily be involved in cardiac phenotypes specifically, but could be relevant to the development of precursor cells or even to broader germ layers such as mesoderm cells (Kiecker et al. 2015).

Another question that may arise in response to the results of the single cell time course project in *Chapter III* is the interpretation of distinct differentiation trajectories and potentially different cell types at the end of the time course. We found that, in later stages of differentiation (days 7, 11, and 15), most cells have either high gene expression of cardiac troponin T (*TNNT2*) and associated genes (such as myosin light chain/*MYL7*), or high gene expression of a collagen-coding gene (*COL3A1*) and associated genes (such as vimentin/*VIM*), as discovered by unsupervised clustering. Cells broadly express either of these gene sets in a mutually exclusive manner, suggesting that these gene sets represent two distinct cell types. Additionally, at these later time points, a much smaller group of cells expresses neither the *TNNT2* or *COL3A1* gene sets at high levels, but instead expresses genes such as *APOA1* and *AFP*. The focus of the single-cell time course project was not to fully characterize these cell types, but instead to distinguish the effect of cell type composition overall from the effect of gene regulation. However, the

identity of these cell types and the circumstances in which each trajectory might be favored is an interesting question.

It is clear by these data that there are differences in gene expression trajectory and ultimate cell fate that arise in response to the same differentiation protocol. What could these cell types be? And what factors would cause a cell line to favor one differentiation trajectory and ultimate cell type at the expense of another? Similar questions have previously been asked by others who perform cardiomyocyte differentiation from embryonic stem cells (D'Antonio-Chronowska et al. 2019). In the referenced study, cell lines undergoing cardiac differentiation resulted in a heterogeneous cell type population, which were identified as either true cardiomyocyte cells which exhibit mechanical beating and have high expression of *TNNT2*, or "epicardium-derived cells" which do not exhibit mechanical beating and have high expression of gene markers such as *VIM* and *TAGLN*. The study demonstrates that these two cell types are present in varying proportions in each of their individual cell lines, and suggests that this cell fate decision can be influenced by genetic factors, such as variability in X chromosome gene dosage (D'Antonio-Chronowska et al. 2019).

As yet, there is no strong evidence to confirm that the non-cardiomyocyte cell type present in my single-cell differentiation time course is related to epicardium-derived cells. However, the cardiomyocyte vs. epicardium framework explored by D'Antonio-Chronowska et al. may be useful in understanding the distinct differentiation trajectories present in our cardiac differentiations as well; specifically, that samples may progress down one differentiation pathway over another in part due to genetic factors. In the case of the terminal non-cardiomyocyte cells expressing *COL3A1* in my samples, I hypothesize that these cells may represent an endothelial or cardiac fibroblast cell type. Cardiac fibroblasts arise from the

epicardium cell lineage, and express gene markers also found in my data, such as collagen and vimentin (Brade et al. 2013, Ieda et al. 2009, Zhang et al. 2019). The gene expression profile of these *COL3A1*-expressing cells, which includes high expression of genes related to extracellular matrix and physical cellular structure, implies that these terminal cells may be involved in providing some kind of structural support, perhaps as accessories to true beating cardiomyocytes.

To determine whether differentiation trajectory and ultimate cell fate decision is influenced by genetic factors, it may be useful to perform cardiomyocyte differentiation on the same cell lines from the same individuals in multiple replicates, and compare the differentiation trajectories between these replicates. Differentiation using the same individuals was performed in my study at least twice (during the collection of bulk RNA-seq and during the collection of single-cell RNA-seq data), and the relatively high correlation between these two replicates suggests that there may be genetic factors involved in this trajectory decision -- although more rigorous testing must be performed to make any conclusions about this claim. Another way to gain insight into this question is to investigate whether there are systematic differences between cell lines even as induced pluripotent stem cells (Day 0) before any differentiation protocol is performed, and whether these differences correlate with the ultimate trajectory of these cell lines during differentiation. Recent studies have suggested that there may be genes whose expression level at the iPSC stage correlates with downstream differentiation efficiency in a predictable manner (Cuomo et al. 2020). Their results suggest that the decision for ultimate cell type trajectories remains consistent within a cell line, and that iPSCs from those cell lines exhibit distinct gene expression profiles that can be used to accurately predict their differentiation trajectories even before differentiation begins. This is an intriguing possibility, and more work should be performed to investigate whether the cell lines used in my studies also exhibit distinct

gene expression profiles early on that may correlate with the outcomes of any subsequent differentiation.

In addition to those outlined above, this work could be followed up with several other future directions. One could use the same cell lines used for the cardiomyocyte differentiation to perform targeted differentiation to other terminal cell types, both within the mesoderm lineage and outside of it. Dynamic eQTLs could be identified in the differentiation of other terminal cell types, and these effects could be compared with those found in the cardiomyocyte differentiation time course. It is possible that many of these dynamic genetic effects are specific to each differentiation and the terminal cell type that arises from it, but some dynamic genetic effects may be shared between distinct differentiation protocols and end points (Dimas et al. 2009). In particular, dynamic effects that arise during the intermediate stages of the cardiomyocyte time course, and that are replicated in other differentiation time courses, may be effects related to the process of differentiation overall, rather than specific to the differentiation of one terminal cell type. Some dynamic genetic effects may be shared between cell types that arise from one germ layer, such as mesoderm, but not shared in cell types from either endoderm or ectoderm (Hutchins et al. 2017). A project that investigates the presence of dynamic eQTLs across the differentiation of all germ layers and many terminal cell types could provide insight into how and when these dynamic effects function during human development more broadly.

Along with identifying dynamic gene regulatory effects in multiple cell types, we can also identify these dynamic effects in response to various environmental perturbations or stress conditions. In a recent study, iPSC-derived cardiomyocyte cells were subjected to oxygen deprivation, and gene expression was measured at several time points before, during, and after this hypoxic stress (Ward et al., n.d.). The authors were able to identify dynamic genetic effects

associated with gene expression changes that occur as a response to hypoxic stress over time. Similar studies can be performed to identify dynamic genetic effects as a function of differentiation stage, stress response, or a combination of the two. Cell type-relevant stressful stimuli such as oxygen deprivation may be applied at various time points over the course of a differentiation from iPSC to cardiomyocyte cells, for many individuals. We may investigate whether the dynamic effect of gene regulation on gene expression differs between individuals due to hypoxic stress in a manner that depends on differentiation stage. For example, we might see the dynamic stress response persist over time along the differentiation, or be washed away by the time the cells have fully differentiated into their terminal cell type. We could also ask whether certain stages of a differentiation are more sensitive or more robust in their dynamic response to environmental stress, and whether this sensitivity or robustness is genetically encoded such that it differs between individuals.

Finally, we can use the differentiation time course study design, particularly with single cell RNA sequencing, to investigate gene expression variance between individuals, between differentiation stages, and between single cells. The regulation of gene expression variance is one mechanism by which a biological system can maintain stable function despite external perturbations or stochastic noise. To maintain the fidelity of cellular behavior, the genome may be under regulatory pressure to minimize the variance, and thus maximize robustness, of a phenotype. Genetic regulation is required to be robust in order to maintain the phenotype or identity of a cell, which may be especially important during a process such as cell type differentiation. For example, the gene expression variance of pluripotency factors Oct4 and Nanog in a population of embryonic stem cells has been linked to the differentiation potential of individual cells, with the expression variance providing a dynamic range that contributes to

phenotypic robustness (Kalmar et al. 2009; Hough et al. 2009). Many dynamic physiological processes must also be robust, and loss of robustness is associated with certain clinically relevant phenotypes and complex genetic disease (Gibson et al. 2009, Ogbunugafor et al. 2010). Assessing inter-individual variation of robustness and gene expression variance in differentiating cardiomyocytes may improve likelihood of detection of cardiovascular disease-related genetic variants. Furthermore, assessing variation in cell-to-cell gene expression variance between differentiation stages may enhance our understanding of the temporal dynamics of the tight regulation of gene expression, or the relaxation of these constraints, across development.

Conclusion

The studies detailed in this thesis demonstrate how a differentiation time course can be used to investigate the dynamics of gene regulation and cell type composition changes over time. Dynamic gene regulatory loci can reveal transient effects not found in mature tissues, and demonstrate the advantages of using a differentiation time course study design. Samples undergoing differentiation may simultaneously experience changes in their gene regulatory profiles as well as their identities and overall cell composition, and these factors can be disentangled with the use of single-cell RNA sequencing. Identifying and separating distinct differentiation trajectories within a heterogeneous population may also uncover relevant dynamic genetic effects that were previously obscured. These dynamic genetic effects may provide new insight into the genetic architecture underlying human development, complex phenotypes, and disease.

References

Abbott L, Bryant S, Churchhouse C, Ganna A, Howrigan D, Palmer D, Neale B, Walters R, Carey C, The Hail Team, V. Anttila, K. Aragam, A. Baumann, J. Cole, M. J. Daly, R. Damian, M. Haas, J. Hirschhorn, Er. Jones, R. Munshi, M. Rivas, S. Vedantam. (2018). UK Biobank–Neale lab; http://www.nealelab.is/uk-biobank/.

Ahmad, F., Banerjee, S. K., Lage, M. L., Huang, X. N., Smith, S. H., Saba, S., Rager, J., Conner, D. A., Janczewski, A. M., Tobita, K., Tinney, J. P., Moskowitz, I. P., Perez-Atayde, A. R., Keller, B. B., Mathier, M. A., Shroff, S. G., Seidman, C. E., & Seidman, J. G. (2008). The role of cardiac troponin T quantity and function in cardiac development and dilated cardiomyopathy. PloS One, 3(7), e2642.

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nature Reviews. Genetics, 16(4), 197–212.

Alberts, B., Bray, D., J. H. W., Hunt, Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1989). Molecular Biology of the Cell. Courier Corporation.

Arking, D. E., Pulit, S. L., Crotti, L., van der Harst, P., Munroe, P. B., Koopmann, T. T., Sotoodehnia, N., Rossin, E. J., Morley, M., Wang, X., Johnson, A. D., Lundby, A., Gudbjartsson, D. F., Noseworthy, P. A., Eijgelsheim, M., Bradford, Y., Tarasov, K. V., Dörr, M., Müller-Nurasyid, M., … Newton-Cheh, C. (2014). Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nature Genetics, 46(8), 826–836.

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., Lambourne, J. J., Sivapalaratnam, S., Downes, K., Kundu, K., Bomba, L., Berentsen, K., Bradley, J. R., Daugherty, L. C., Delaneau, O., … Soranzo, N. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell, 167(5), 1415–1429.e19.

Banovich, N. E., Li, Y. I., Raj, A., Ward, M. C., Greenside, P., Calderon, D., Tung, P. Y., Burnett, J. E., Myrthil, M., Thomas, S. M., Burrows, C. K., Romero, I. G., Pavlovic, B. J., Kundaje, A., Pritchard, J. K., & Gilad, Y. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. Genome Research, 28(1), 122–131.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., & Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research, 24(1), 14–24.

Bis, J. C., Kavousi, M., Franceschini, N., Isaacs, A., Abecasis, G. R., Schminke, U., Post, W. S., Smith, A. V., Cupples, L. A., Markus, H. S., Schmidt, R., Huffman, J. E., Lehtimäki, T., Baumert, J., Münzel, T., Heckbert, S. R., Dehghan, A., North, K., Oostra, B., … CARDIoGRAM Consortium. (2011). Meta-analysis of genome-wide association studies

from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. Nature Genetics, 43(10), 940–947.

Bizy, A., Guerrero-Serna, G., Hu, B., Ponce-Balbuena, D., Willis, B. C., Zarzoso, M., Ramirez, R. J., Sener, M. F., Mundada, L. V., Klos, M., Devaney, E. J., Vikstrom, K. L., Herron, T. J., & Jalife, J. (2013). Myosin light chain 2-based selection of human iPSC-derived early ventricular cardiac myocytes. Stem Cell Research, 11(3), 1335–1347.

Brade, T., Pane, L. S., Moretti, A., Chien, K. R., & Laugwitz, K.-L. (2013). Embryonic heart progenitors and cardiogenesis. Cold Spring Harbor Perspectives in Medicine, 3(10), a013847.

Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: a theory. Science, 165(3891), 349–357.

Burke, M. A., Cook, S. A., Seidman, J. G., & Seidman, C. E. (2016). Clinical and Mechanistic Insights Into the Genetics of Cardiomyopathy. Journal of the American College of Cardiology, 68(25), 2871–2886.

Burridge, P. W., Matsa, E., Shukla, P., Lin, Z. C., Churko, J. M., Ebert, A. D., Lan, F., Diecke, S., Huber, B., Mordwinkin, N. M., Plews, J. R., Abilez, O. J., Cui, B., Gold, J. D., & Wu, J. C. (2014). Chemically defined generation of human cardiomyocytes. Nature Methods, 11(8), 855–860.

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B. A., Stirrups, K., König, I. R., Cazier, J.-B., Johansson, A., Hall, A. S., Lee, J.-Y., Willer, C. J., Chambers, J. C., … Samani, N. J. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nature Genetics, 45(1), 25–33.

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

Cuomo, A. S. E., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K. N., HipSci Consortium, Vallier, L., Marioni, J. C., Chhatriwala, M., & Stegle, O. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nature Communications, 11(1), 810.

Dambrot, C., Passier, R., Atsma, D., & Mummery, C. L. (2011). Cardiomyocyte differentiation of pluripotent stem cells and their use as cardiac disease models. Biochemical Journal, 434(1), 25–35.

D’Antonio-Chronowska, A., Donovan, M. K. R., Young Greenwald, W. W., Nguyen, J. P., Fujita, K., Hashem, S., Matsui, H., Soncin, F., Parast, M., Ward, M. C., Coulet, F., Smith, E. N., Adler, E., D’Antonio, M., & Frazer, K. A. (2019). Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. Stem Cell Reports, 13(5), 924–938.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., & Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature, 482(7385), 390–394.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J., Deloukas, P., Dermitzakis, E. T., & Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science, 325(5945), 1246–1250.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics , 29(1), 15–21.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. American Journal of Human Genetics, 93(5), 779–797.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews. Genetics, 11(6), 446–450.

Elshourbagy, N. A., Boguski, M. S., Liao, W. S., Jefferson, L. S., Gordon, J. I., & Taylor, J. M. (1985). Expression of rat apolipoprotein A-IV and A-I genes: mRNA induction during development and in response to glucocorticoids and insulin. Proceedings of the National Academy of Sciences of the United States of America, 82(23), 8242–8246.

Ernst, J., & Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. In Nature Protocols (Vol. 12, Issue 12, pp. 2478–2492). https://doi.org/10.1038/nprot.2017.124

Gamazon, E. R., Huang, R. S., Dolan, M. E., Cox, N. J., & Im, H. K. (2012). Integrative genomics: quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. Frontiers in Genetics, 3, 202.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell, 176(6), 1516.

Gibson, G. (2009). Decanalization and the origin of complex disease. Nature Reviews. Genetics, 10(2), 134–140.

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. In Nature (Vol. 550, Issue 7675, pp. 204–213). https://doi.org/10.1038/nature24277

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S. R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D. A., Rozenblatt-Rosen, O., Zhang, F., &

Regev, A. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nature Methods, 14(10), 955–958.

Hensman J, Matthews A. G. de G., Ghahramani Z. (2015). Scalable variational Gaussian process classification. Proc. Mach. Learn. Res. 38, 351–360.

Heron, M. (2019). Deaths: Leading Causes for 2017. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 68(6), 1–77.

Hong, K.-W., Lim, J. E., Kim, J. W., Tabara, Y., Ueshima, H., Miki, T., Matsuda, F., Cho, Y. S., Kim, Y., & Oh, B. (2014). Identification of three novel genetic variations associated with electrocardiographic traits (QRS duration and PR interval) in East Asians. Human Molecular Genetics, 23(24), 6659–6667.

Hough, S. R., Laslett, A. L., Grimmond, S. B., Kolle, G., & Pera, M. F. (2009). A Continuum of Cell States Spans Pluripotency and Lineage Commitment in Human Embryonic Stem Cells. In PLoS ONE (Vol. 4, Issue 11, p. e7708). https://doi.org/10.1371/journal.pone.0007708

Hutchins, A. P., Yang, Z., Li, Y., He, F., Fu, X., Wang, X., Li, D., Liu, K., He, J., Wang, Y., Chen, J., Esteban, M. A., & Pei, D. (2017). Models of global gene expression define major domains of cell type and tissue identity. Nucleic Acids Research, 45(5), 2354–2367.

Ieda, M., Tsuchihashi, T., Ivey, K. N., Ross, R. S., Hong, T.-T., Shaw, R. M., & Srivastava, D. (2009). Cardiac fibroblasts regulate myocardial proliferation through beta1 integrin signaling. Developmental Cell, 16(2), 233–244.

Joehanes, R., Zhang, X., Huan, T., Yao, C., Ying, S.-X., Nguyen, Q. T., Demirkale, C. Y., Feolo, M. L., Sharopova, N. R., Sturcke, A., Schäffer, A. A., Heard-Costa, N., Chen, H., Liu, P.-C., Wang, R., Woodhouse, K. A., Tanriverdi, K., Freedman, J. E., Raghavachari, N., … Munson, P. J. (2017). Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biology, 18(1), 16.

Josowitz, R., Carvajal-Vergara, X., Lemischka, I. R., & Gelb, B. D. (2011). Induced pluripotent stem cell-derived cardiomyocytes as models for genetic cardiovascular disorders. Current Opinion in Cardiology, 26(3), 223–229.

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., & Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biology, 7(7), e1000149.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., Gate, R. E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L. A., & Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology, 36(1), 89–94.

Kiecker, C., Bates, T., & Bell, E. (2016). Molecular specification of germ layers in vertebrate embryos. Cellular and Molecular Life Sciences: CMLS, 73(5), 923–947.

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Favé, M.-J., Zhu, X., Potash, J. B., Weissman, M. M., Shi, J., Levinson, D. F., Awadalla, P., Mostafavi, S., Montgomery, S. B., & Battle, A. (2017). Allele-specific expression reveals interactions between genetic variation and environment. Nature Methods, 14(7), 699–702.

Kubara, K., Yamazaki, K., Ishihara, Y., Naruto, T., Lin, H.-T., Nishimura, K., Ohtaka, M., Nakanishi, M., Ito, M., Tsukahara, K., Morio, T., Takagi, M., & Otsu, M. (2018). Status of KRAS in iPSCs Impacts upon Self-Renewal and Differentiation Propensity. Stem Cell Reports, 11(2), 380–394.

Lappalainen, T., Sammeth, M., Friedländer, M. R., ’t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506–511.

Lazarevich, N. L. (2000). Molecular mechanisms of alpha-fetoprotein gene expression. Biochemistry. Biokhimiia, 65(1), 117–133.

Lázaro-Gredilla M, Van Vaerenbergh S, Lawrence N. D. (2012). Overlapping mixtures of Gaussian processes for the data association problem. Pattern Recognit. 45, 1386–1395. doi:10.1016/j.patcog.2011.10.004

Lian, X., Zhang, J., Azarin, S. M., Zhu, K., Hazeltine, L. B., Bao, X., Hsiao, C., Kamp, T. J., & Palecek, S. P. (2013). Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/β-catenin signaling under fully defined conditions. Nature Protocols, 8(1), 162–175.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics , 30(7), 923–930.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Systems, 1(6), 417–425.

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. Science, 352(6285), 600–604.

Lu, T.-Y., & Yang, L. (2011). Uses of cardiomyocytes generated from induced pluripotent stem cells. Stem Cell Research & Therapy, 2(6), 44.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell, 161(5), 1202–1214.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747–753.

McInnes, L., Healy, J., Saul, N. & Großberger, L (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw., https://doi.org/10.21105/joss.00861

Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., … Altshuler, D. (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nature Genetics, 41(3), 334–341.

Narsinh, K., Narsinh, K. H., & Wu, J. C. (2011). Derivation of human induced pluripotent stem cells for cardiovascular disease modeling. Circulation Research, 108(9), 1146–1156.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics, 6(4), e1000895.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genetics, 6(4), e1000888.

Ogbunugafor, C. B., Pease, J. B., & Turner, P. E. (2010). On the possible role of robustness in the evolution of infectious diseases. Chaos , 20(2), 026108.

Oh, Y., Wei, H., Ma, D., Sun, X., & Liew, R. (2012). Clinical applications of patient-specific induced pluripotent stem cells in cardiovascular medicine. Heart , 98(6), 443–449.

Okita, K., Ichisaka, T., & Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature, 448(7151), 313–317.

Okita, K., Matsumura, Y., Sato, Y., Okada, A., Morizane, A., Okamoto, S., Hong, H., Nakagawa, M., Tanabe, K., Tezuka, K.-I., Shibata, T., Kunisada, T., Takahashi, M., Takahashi, J., Saji, H., & Yamanaka, S. (2011). A more efficient method to generate integration-free human iPS cells. Nature Methods, 8(5), 409–412.

Pavlovic, B. J., Blake, L. E., Roux, J., Chavarria, C., & Gilad, Y. (2018). A Comparative Assessment of Human and Chimpanzee iPSC-derived Cardiomyocytes with Primary Heart Tissues. Scientific Reports, 8(1), 15312.

Pedregosa F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.

Brucher, M. Perrot, É. Duchesnay. (2011). Learning scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. American Journal of Human Genetics, 94(4), 559–573.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464(7289), 768–772.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317–330.

Rook, M. B., Evers, M. M., Vos, M. A., & Bierhuizen, M. F. A. (2012). Biology of cardiac sodium channel Nav1.5 expression. Cardiovascular Research, 93(1), 12–23.

Selewa, A., Dohn, R., Eckart, H., Lozano, S., Xie, B., Gauchat, E., Elorbany, R., Rhodes, K., Burnett, J., Gilad, Y., Pott, S., & Basu, A. (2020). Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. Scientific Reports, 10(1), 1535.

Shah, S., Henry, A., Roselli, C., Lin, H., Sveinbjörnsson, G., Fatemifar, G., Hedman, Å. K., Wilk, J. B., Morley, M. P., Chaffin, M. D., Helgadottir, A., Verweij, N., Dehghan, A., Almgren, P., Andersson, C., Aragam, K. G., Ärnlöv, J., Backman, J. D., Biggs, M. L., … Lumbers, R. T. (2020). Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. Nature Communications, 11(1), 163.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Research, 27(3), 491–499.

Stephens, M. (2016). False discovery rates: a new deal. In Biostatistics (p. kxw041). https://doi.org/10.1093/biostatistics/kxw041

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P., & Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genetics, 8(4), e1002639.

Strober B., BennyStrobes/ipsc_cardiomyocyte_differentiation: iPSC-cardiomyocyte differentiation scripts, Version 1.0, Zenodo (2019); http://doi.org/10.5281/zenodo.2591584.

Strober B., BennyStrobes/ipsc_eqtl_results: eQTL results from iPSC-cardiomyocyte differentiation project, Version 1.0, Zenodo (2019); http://doi.org/10.5281/zenodo.2591542.

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.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell, 177(7), 1888–1902.e21.

Tayeb K., karltayeb/ipsc_gp_clustering: split-gpm, Version 1.0, Zenodo (2019); http://doi.org/10.5281/zenodo.2590826.

Tohyama, S., Hattori, F., Sano, M., Hishiki, T., Nagahata, Y., Matsuura, T., Hashimoto, H., Suzuki, T., Yamashita, H., Satoh, Y., Egashira, T., Seki, T., Muraoka, N., Yamakawa, H., Ohgino, Y., Tanaka, T., Yoichi, M., Yuasa, S., Murata, M., … Fukuda, K. (2013). Distinct metabolic flow enables large-scale purification of mouse and human pluripotent stem cell-derived cardiomyocytes. Cell Stem Cell, 12(1), 127–137.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Research, 25(10), 1491–1498.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., & Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature, 509(7500), 371–375.

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature Methods, 12(11), 1061–1063.

van der Wijst, M. G. P., de Vries, D. H., Brugge, H., Westra, H.-J., & Franke, L. (2018). An integrative approach for building personalized gene regulatory networks for precision medicine. Genome Medicine, 10(1), 96.

Ward, M. C., Banovich, N. E., Sarkar, A., Stephens, M., & Gilad, Y. (n.d.). Dynamic effects of genetic variation on gene expression revealed following hypoxic stress in cardiomyocytes. https://doi.org/10.1101/2020.03.28.012823

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), 661–678.

Wen, X., Pique-Regi, R., & Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genetics, 13(3), e1006646.

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biology, 19(1), 15.

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology, 20(1), 59.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., & Tang, F. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature Structural & Molecular Biology, 20(9), 1131–1139.

Yazawa, M., Hsueh, B., Jia, X., Pasca, A. M., Bernstein, J. A., Hallmayer, J., & Dolmetsch, R. E. (2011). Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome. Nature, 471(7337), 230–234.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., & Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science, 318(5858), 1917–1920.

Zeitlinger, J., & Stark, A. (2010). Developmental gene regulation in the era of genomics. Developmental Biology, 339(2), 230–239.

Zhang, J., Tao, R., Campbell, K. F., Carvalho, J. L., Ruiz, E. C., Kim, G. C., Schmuck, E. G., Raval, A. N., da Rocha, A. M., Herron, T. J., Jalife, J., Thomson, J. A., & Kamp, T. J. (2019). Functional cardiac fibroblasts derived from human pluripotent stem cells via second heart field progenitors. Nature Communications, 10(1), 2238.

Zhi, D., Irvin, M. R., Gu, C. C., Stoddard, A. J., Lorier, R., Matter, A., Rao, D. C., Srinivasasainagendra, V., Tiwari, H. K., Turner, A., Broeckel, U., & Arnett, D. K. (2012). Whole-exome sequencing and an iPSC-derived cardiomyocyte model provides a powerful platform for gene discovery in left ventricular hypertrophy. Frontiers in Genetics, 3, 92.

Zhou, Z., Yu, H., Wang, Y., Guo, Q., Wang, L., & Zhang, H. (2016). ZNF606 interacts with Sox9 to regulate chondrocyte differentiation. Biochemical and Biophysical Research Communications, 479(4), 920–926.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics, 48(5), 481–487.

Appendix A: Supplementary Figures and Tables

Supplementary Figures for Chapter II

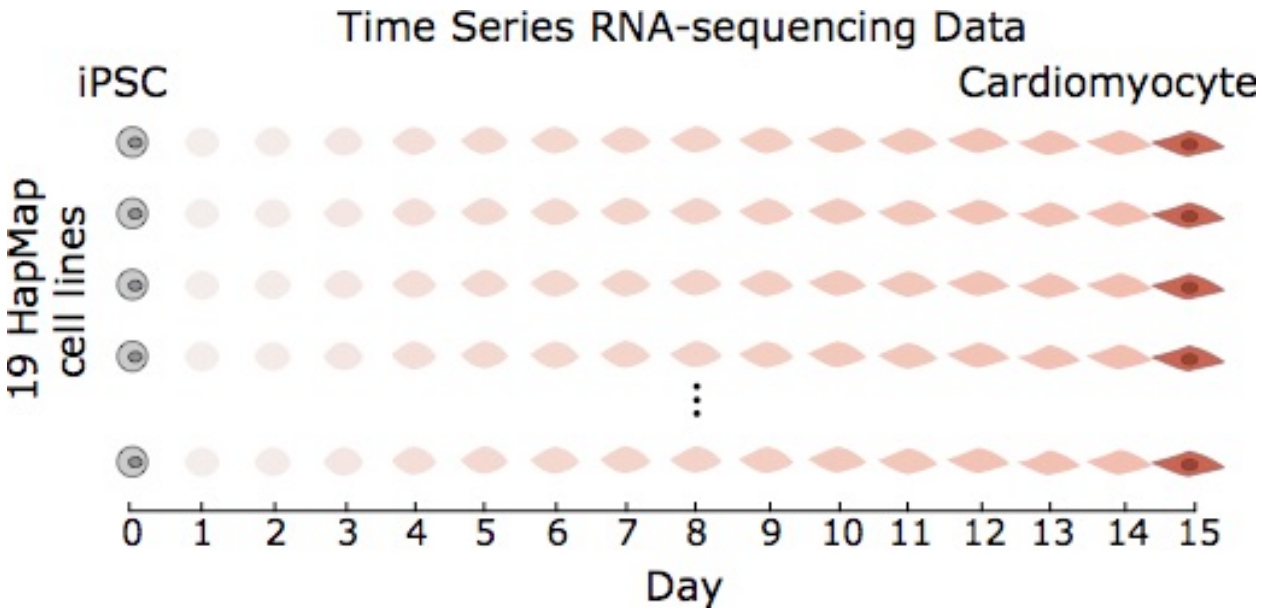


Fig. S2-1. RNA-seq sample collection. Overview of RNA-seq sample collection. In 19 Yoruba HapMap cell lines, RNA was extracted and sequenced every 24 hours at 16 time points, generating 297 RNA-seq samples.

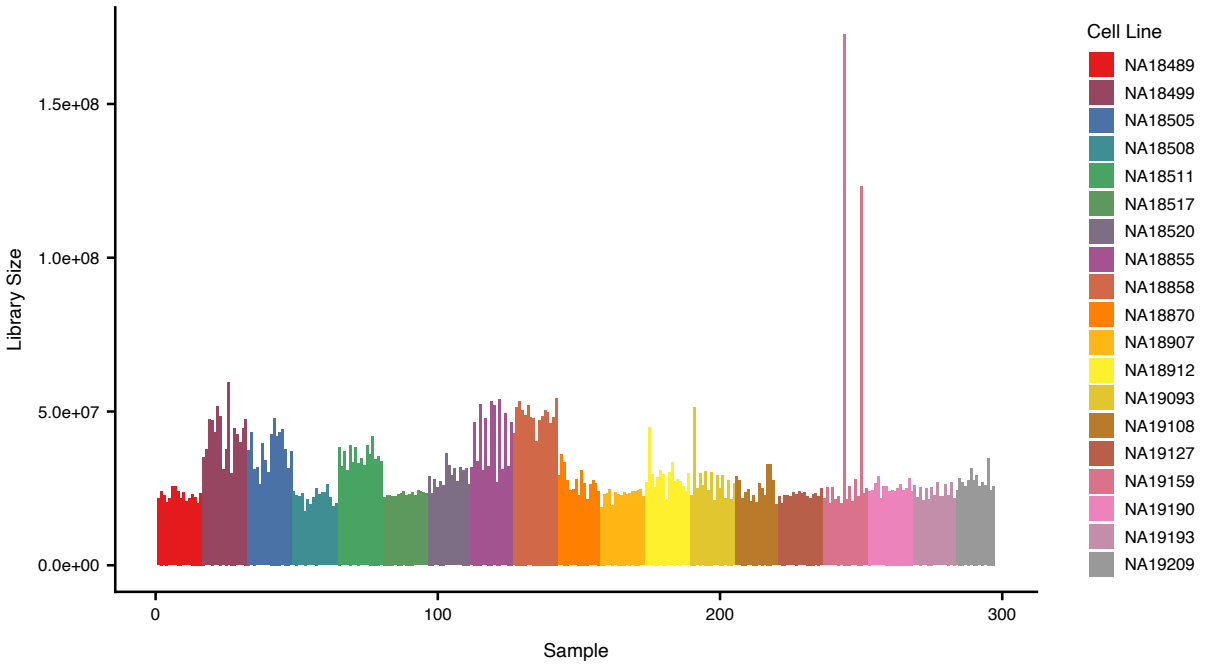


Fig. S2-2. Library size of RNA-seq samples. The library sizes of 297 RNA-seq samples colored by their cell line identity. Within each cell line, samples are ordered along the x-axis by their differentiation time point from day 0 to 15.

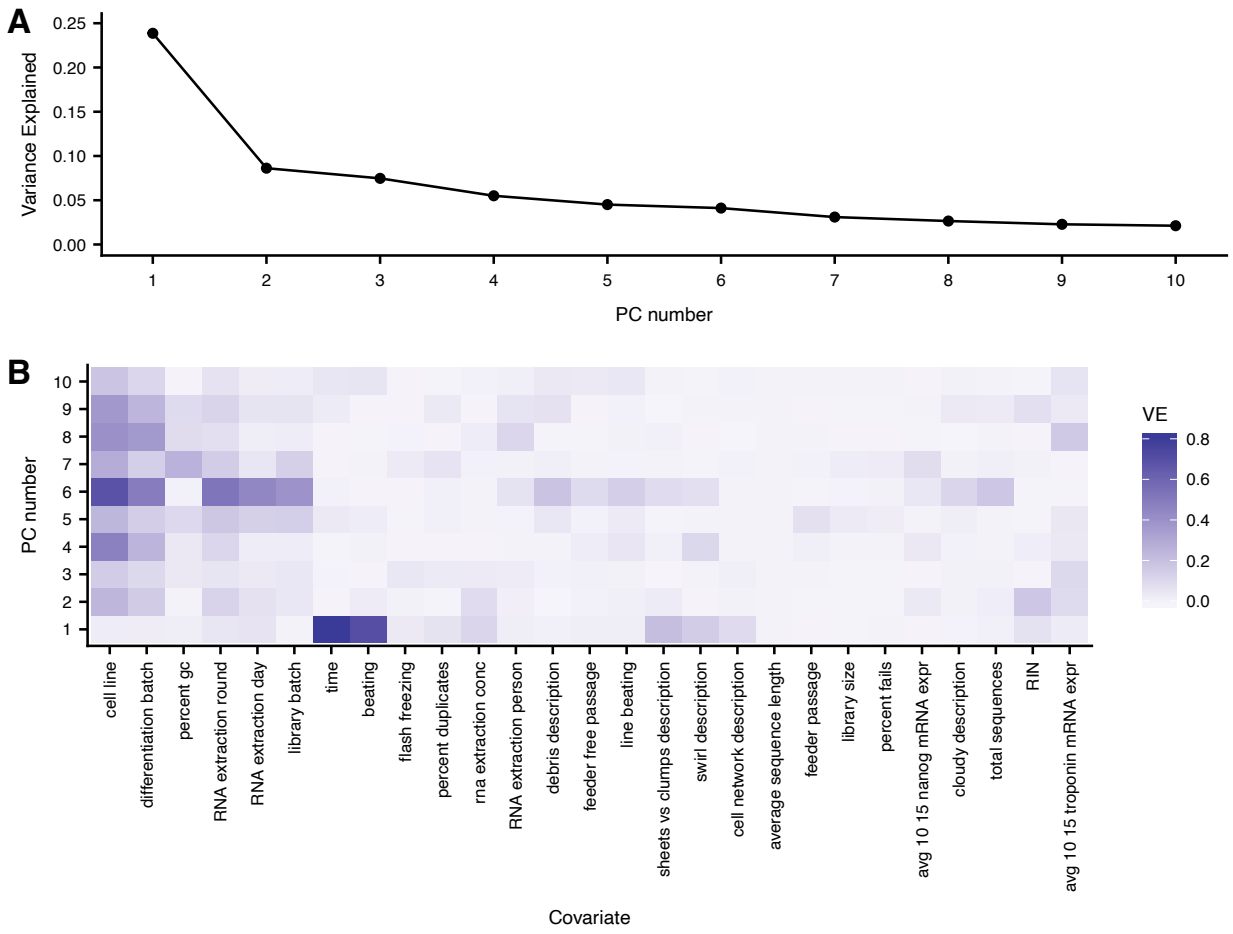


Fig. S2-3. Explaining principal components with sample covariates. (A) Variance in gene expression explained by first 10 gene expression principal components. (B) Variance explained of each gene expression principal component using sample covariates. Adjusted R^2 was used to handle categorical sample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

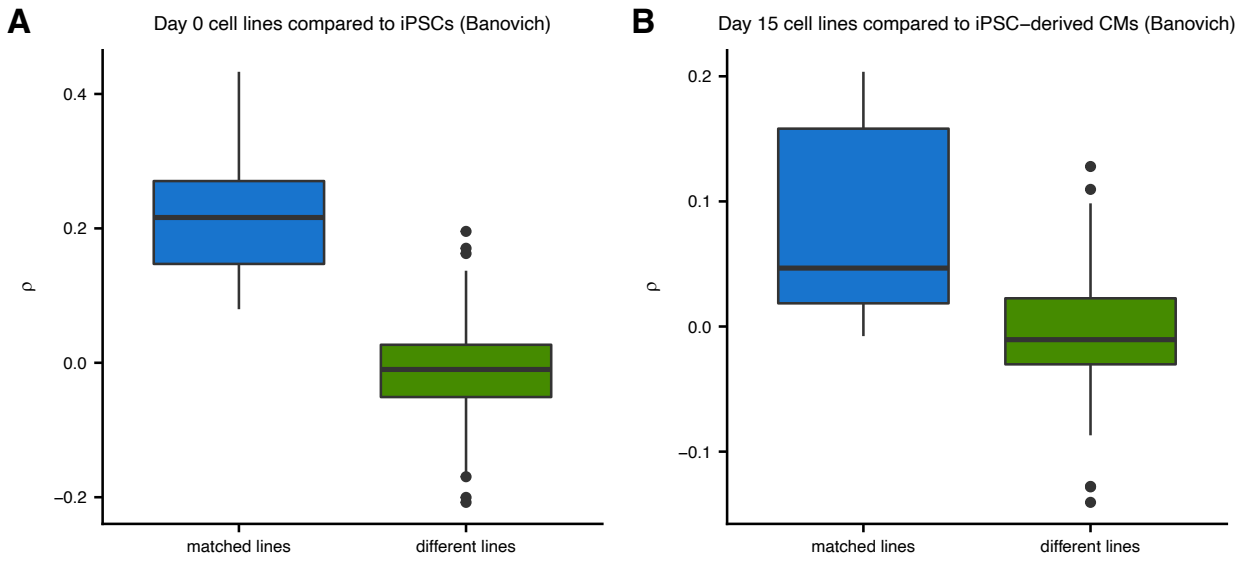


Fig. S2-4. Biological replication of day 0 and day 15 cells. We compared day 0 and day 15 cell lines with matched iPSC lines and iPSC-derived cardiomyocyte lines, respectively, from Banovich et al. (*9*). This analysis was restricted to cell lines present in both data sets. Spearman correlation across 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).

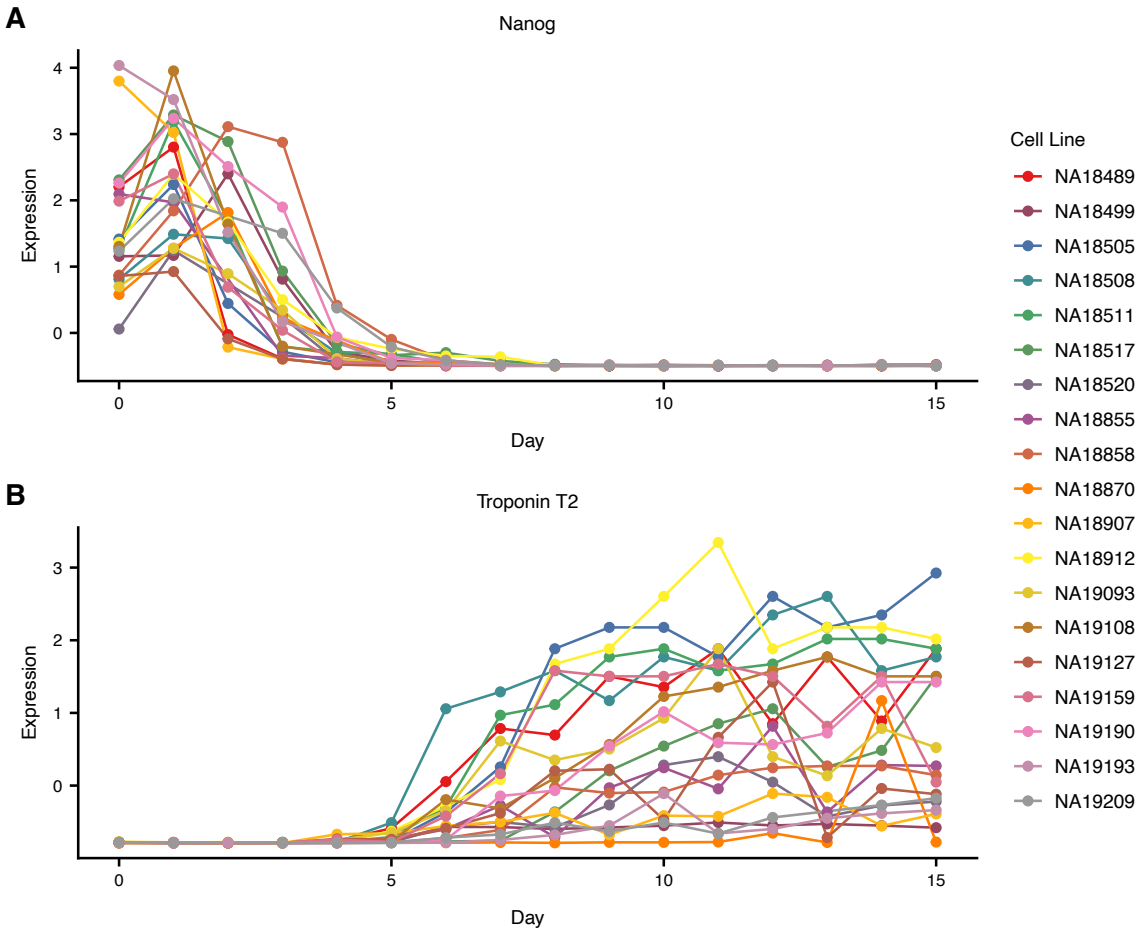


Fig. S2-5. Expression time course of known cell type specific marker genes. Standardized gene expression levels of *Nanog* (A, stem cell marker gene) and *Troponin T2* (B, cardiomyocyte marker gene) across 16 time points (x-axis) and 19 cell lines (colors).

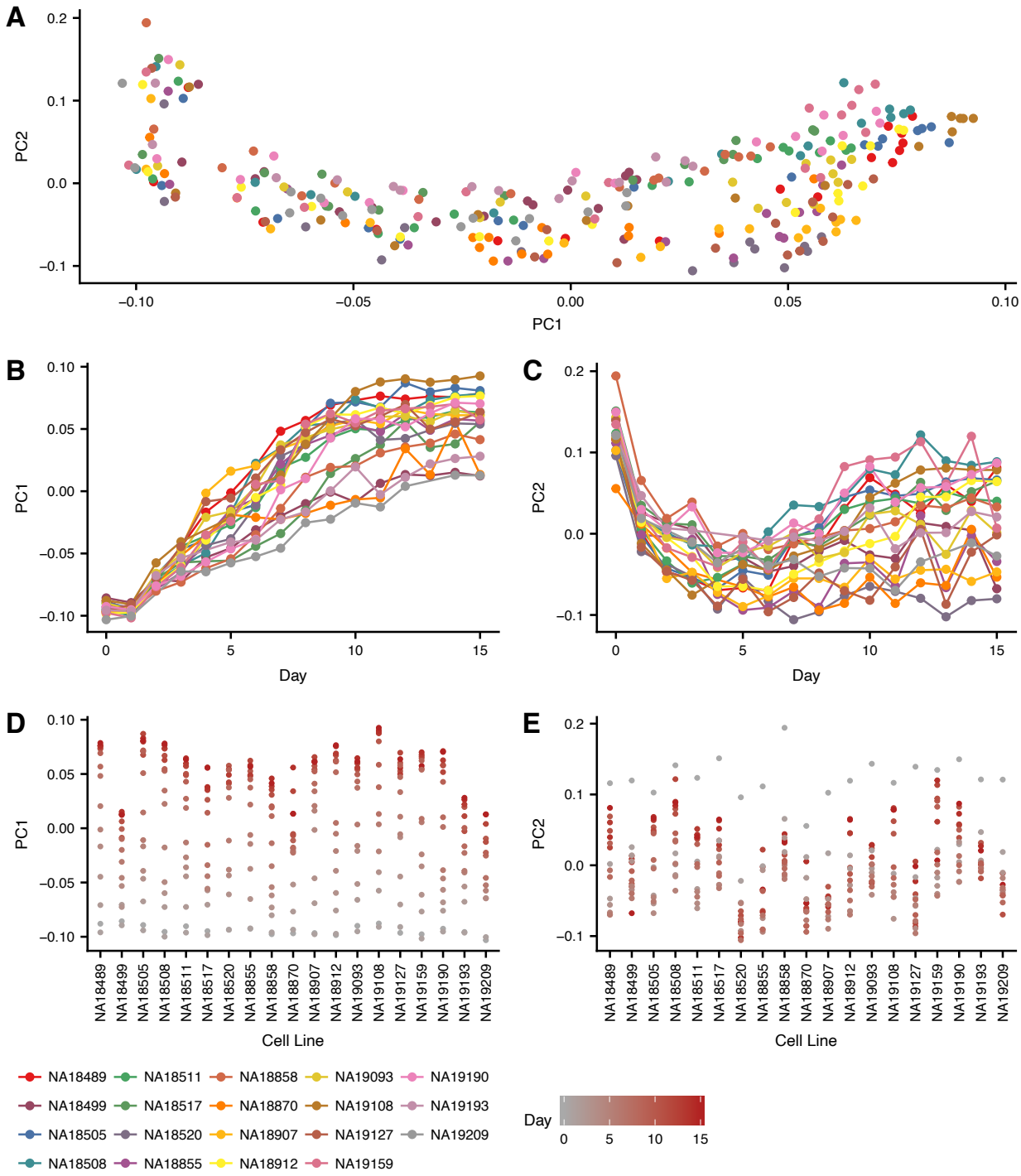


Fig. S2-6. Principal component analysis separated by cell line identity. (A) First two gene expression principal component loadings for all 297 RNA-seq samples, where each sample is colored by its cell line identity. (B, C) Principal component 1 and 2 loadings across 16 time points (x-axis) and 19 cell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

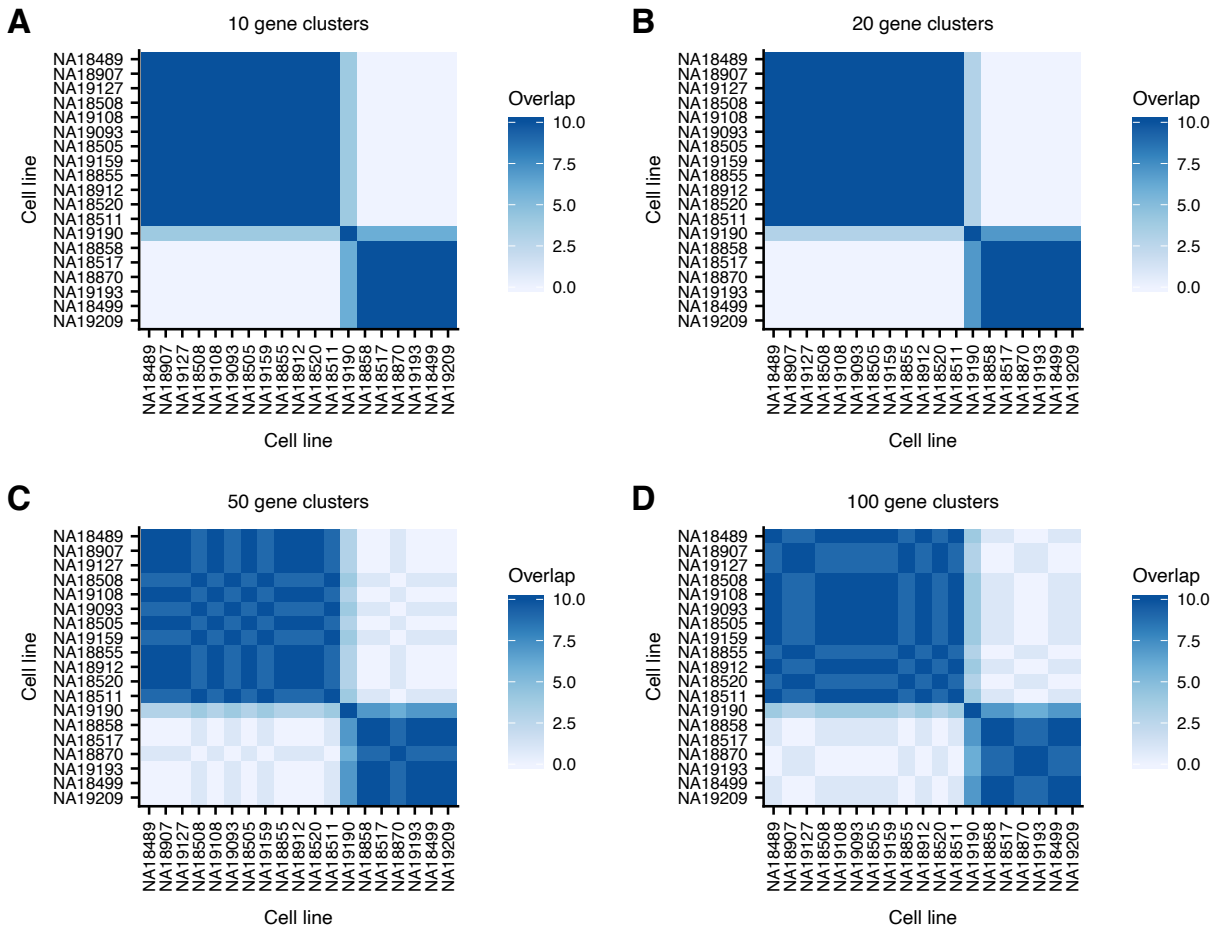


Fig. S2-7. split-GPM cell line cluster assignment robust to hyper-parameter choice. Number of times (out of 10 split-GPM runs with independent, random initializations) that each cell line pair was assigned to the same cell line cluster when 10 (A), 20 (B), 50 (C), and 100 (D) gene clusters were used.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

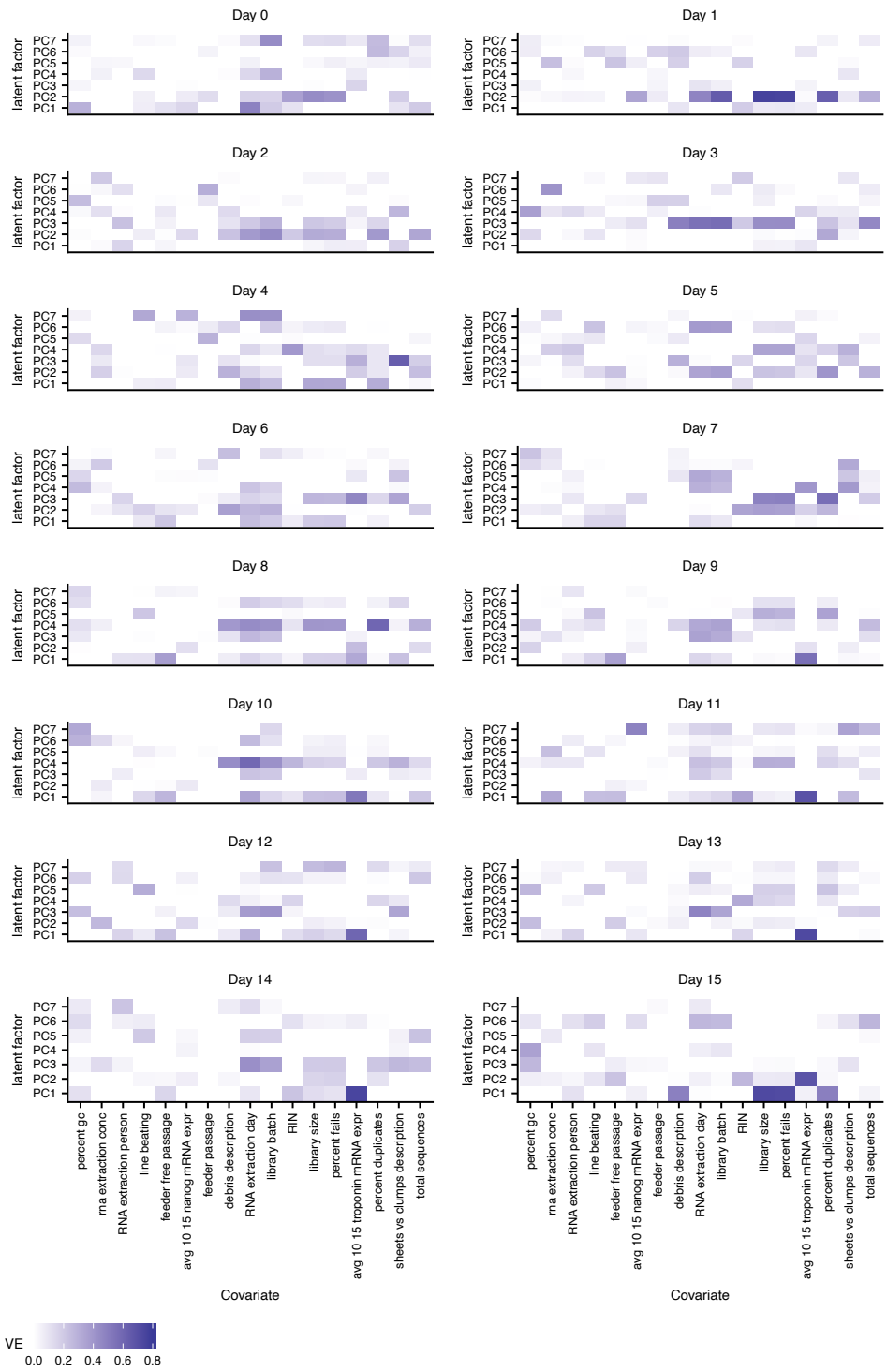


Fig. S2-8. Explaining time step principal components with sample covariates. In each time point independently, variance explained of each raw read count expression principal components (from samples belonging to the corresponding time point) using sample covariates. Adjusted R^2 was used to handle categorical 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.

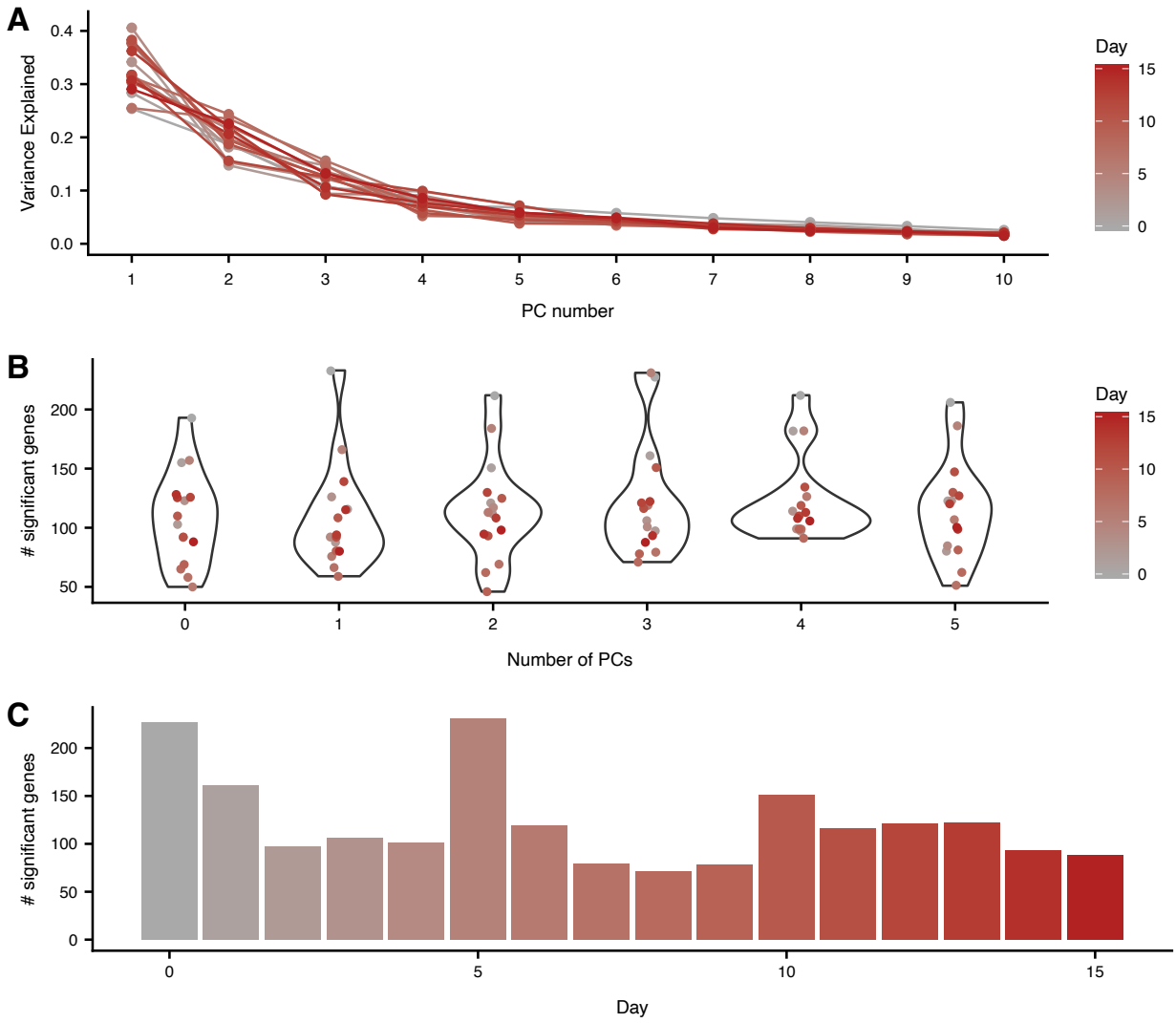


Fig. S2-9. Number of genes with non-dynamic eQTLs. (A) Variance explained of gene expression from samples belonging to a particular time point (color) by the first 10 gene expression PCs (x-axis) computed on samples belonging to that time point. (B) The number of genes with a significant eQTL (eFDR 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

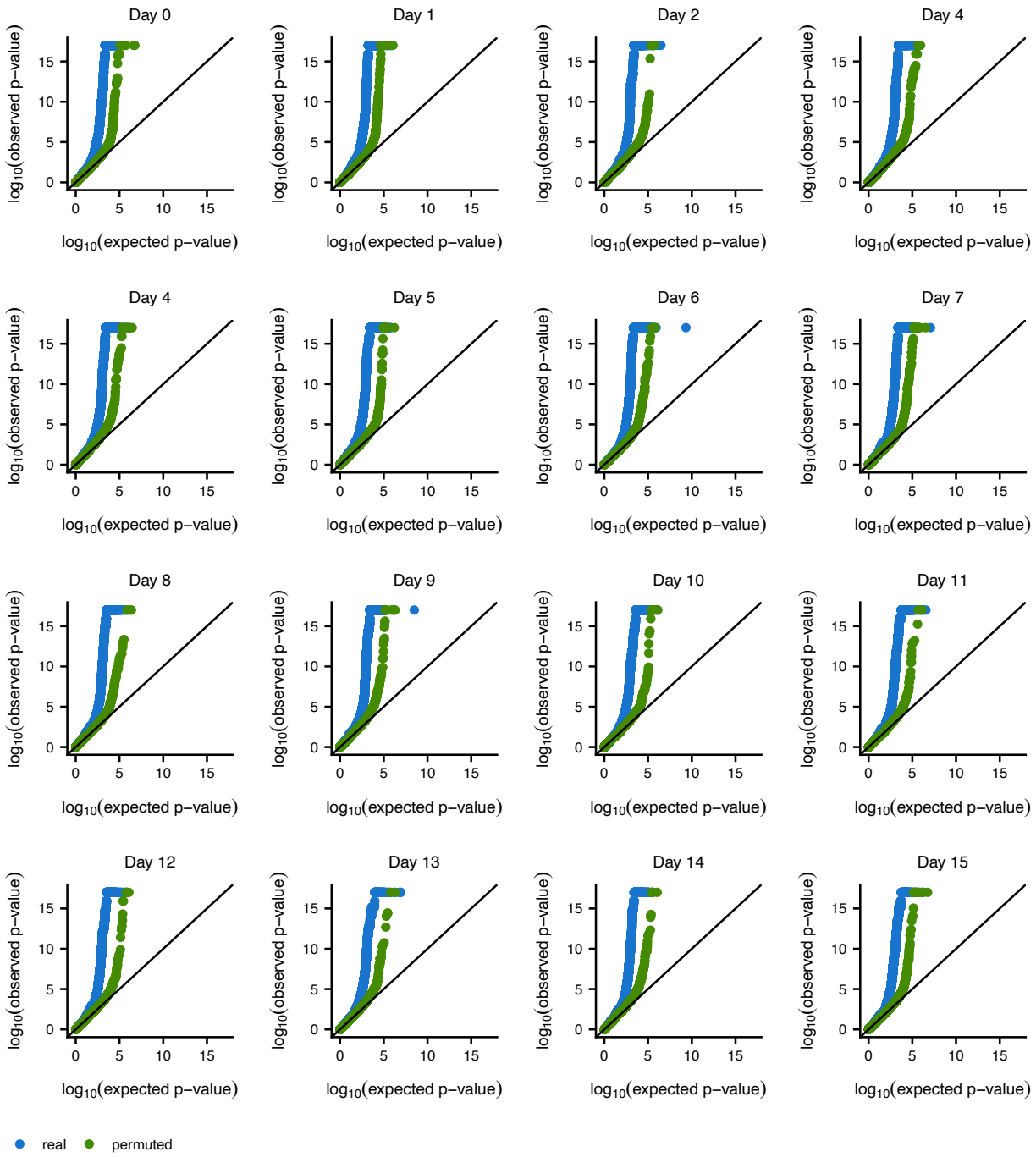


Fig. S2-10. Q-Q plots for non-dynamic eQTLs. Q-Q plot for non-dynamic eQTLs in all 16 time steps. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data (using WASP's permutation strategy) relative to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

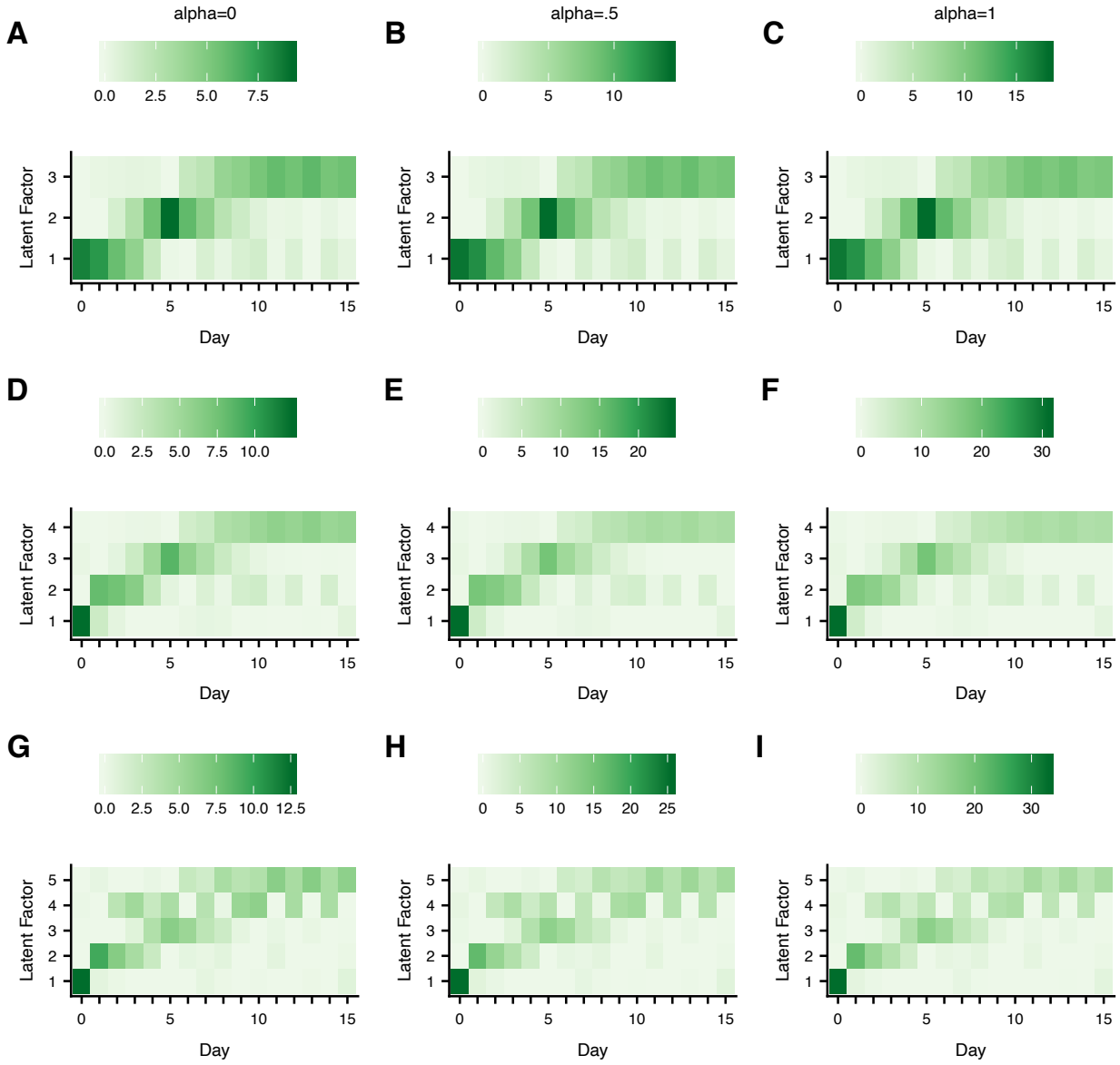


Fig. S2-11. Matrix factorization of eQTL summary statistics. Latent factors identified via sparse non-negative matrix factorization of non-dynamic eQTL -$\log_{10}$ p-values shown for a range of sparse prior choices (alpha; columns) when using 3, 4, and 5 factors (rows).

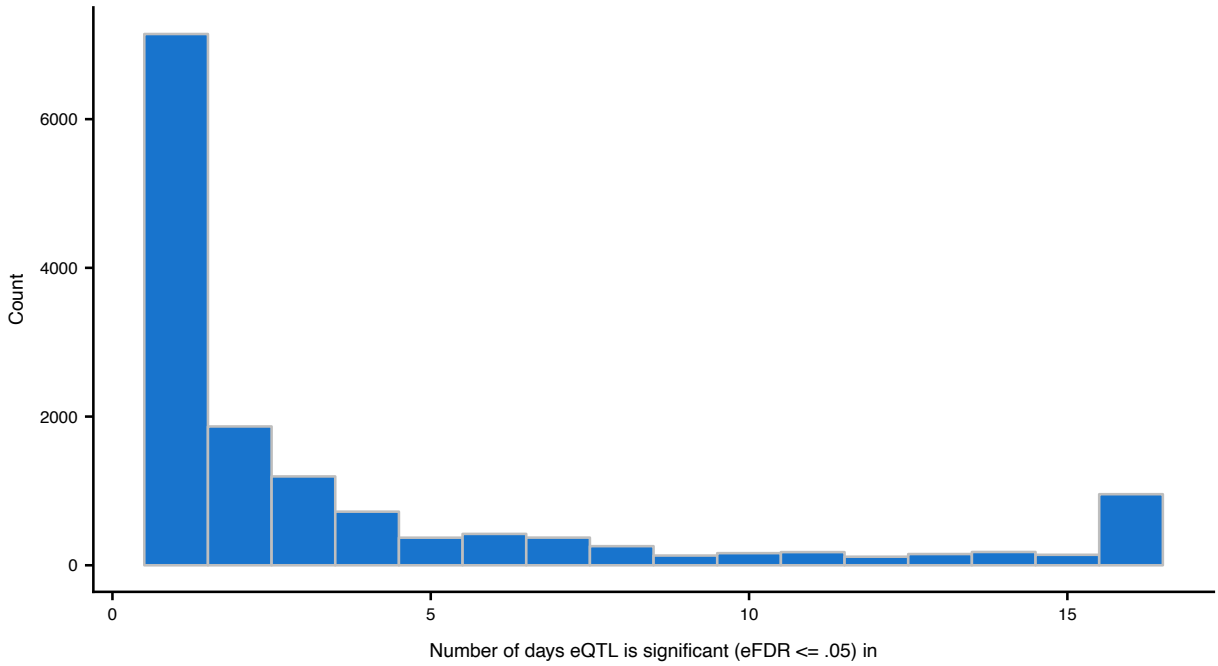


Fig. S2-12. eQTL sharing across time points. The number of days in which each non-dynamic eQTL is significant (eFDR <= .05) for all variant-gene pairs that are significant in at least one day.

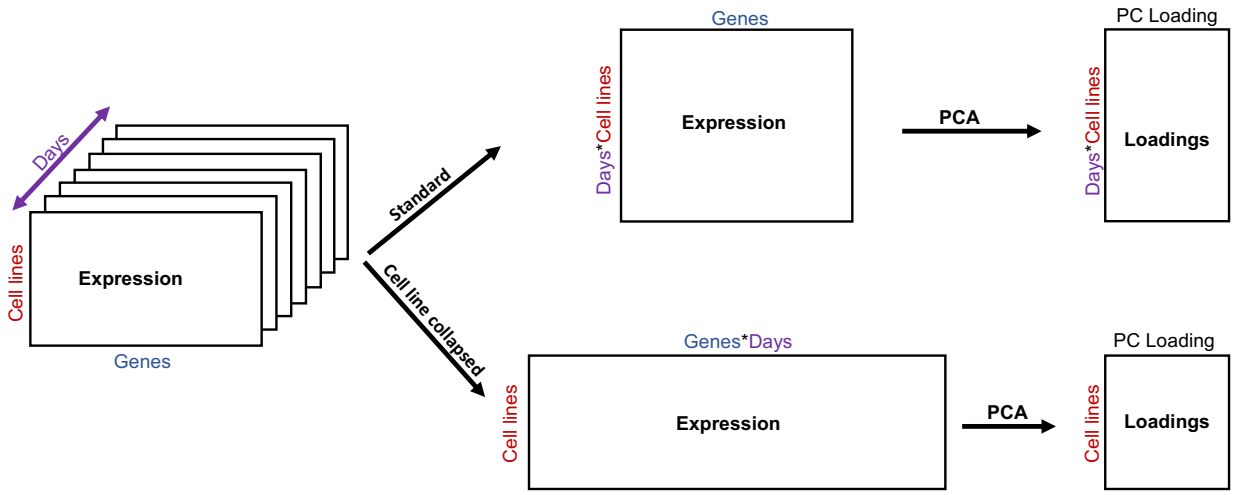


Fig. S2-13. Overview of cell line collapsed PCA. Gene expression can be represented as a three-dimensional matrix spanning days, cell lines, and genes. For standard PCA (top row), we rearrange this gene expression matrix such that rows now correspond to cell lines at specific days (e.g., RNA-seq samples)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

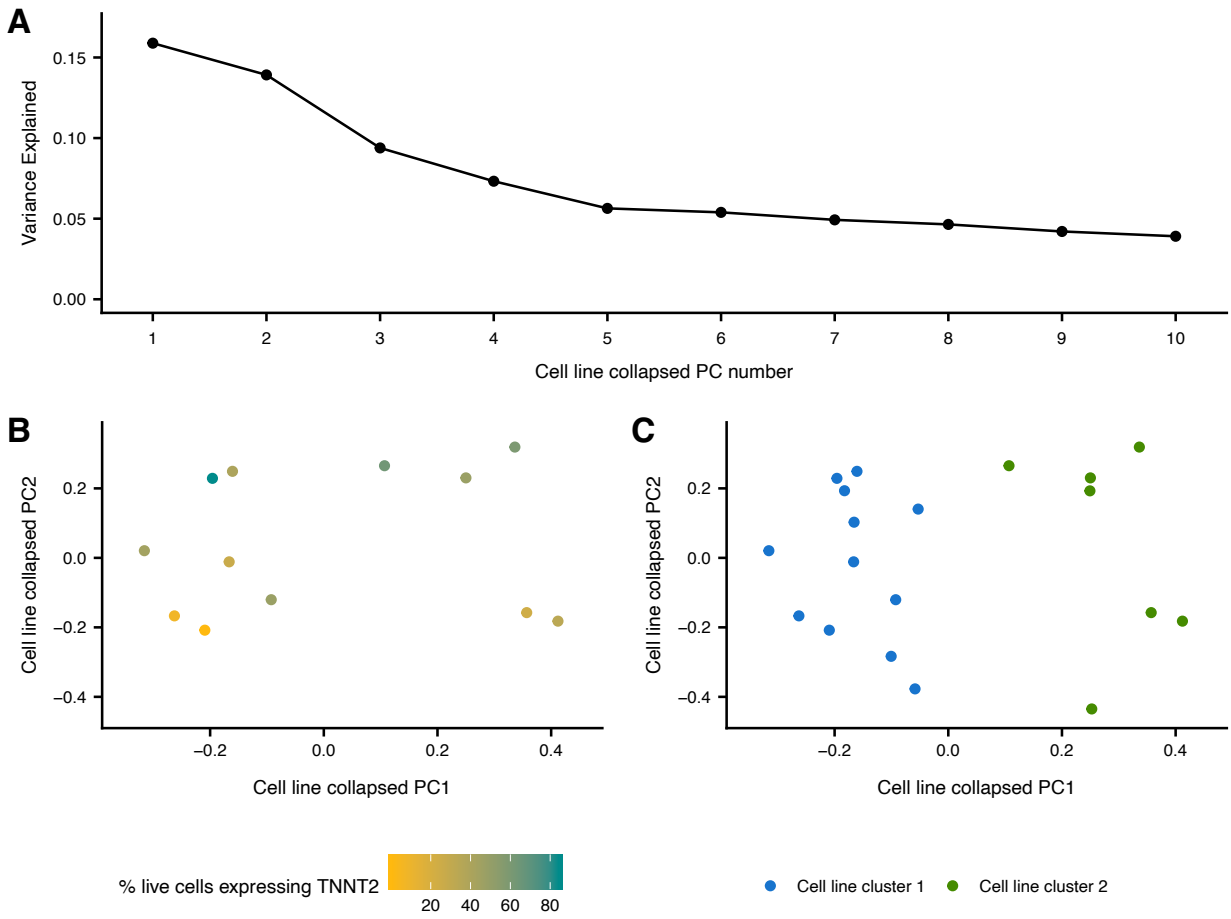


Fig. S2-14. Analysis of cell line collapsed PCs. (A) Variance explained of gene expression by first 10 cell line collapsed principal components. (B, C) First two cell line collapsed principal components where each data point is a cell line colored by its (B) percentage of live cells expressing TNNT2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

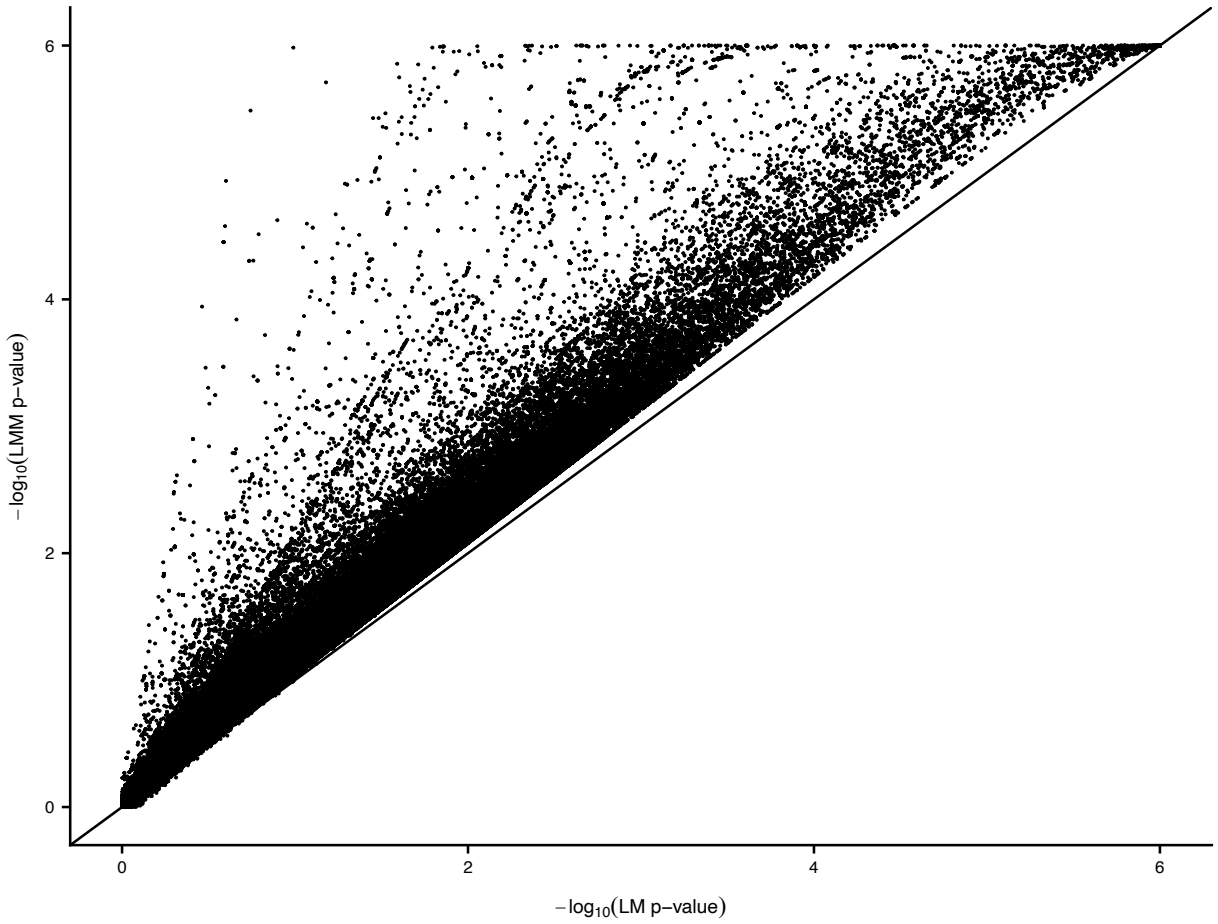


Fig. S2-15. Detecting dynamic eQTLs with gaussian linear mixed model. Comparison of linear dynamic eQTL p-values between gaussian linear model (x-axis) and gaussian linear mixed model with cell line specific random effect (y-axis) across all tested variant-gene pairs (Pearson correlation=.983).

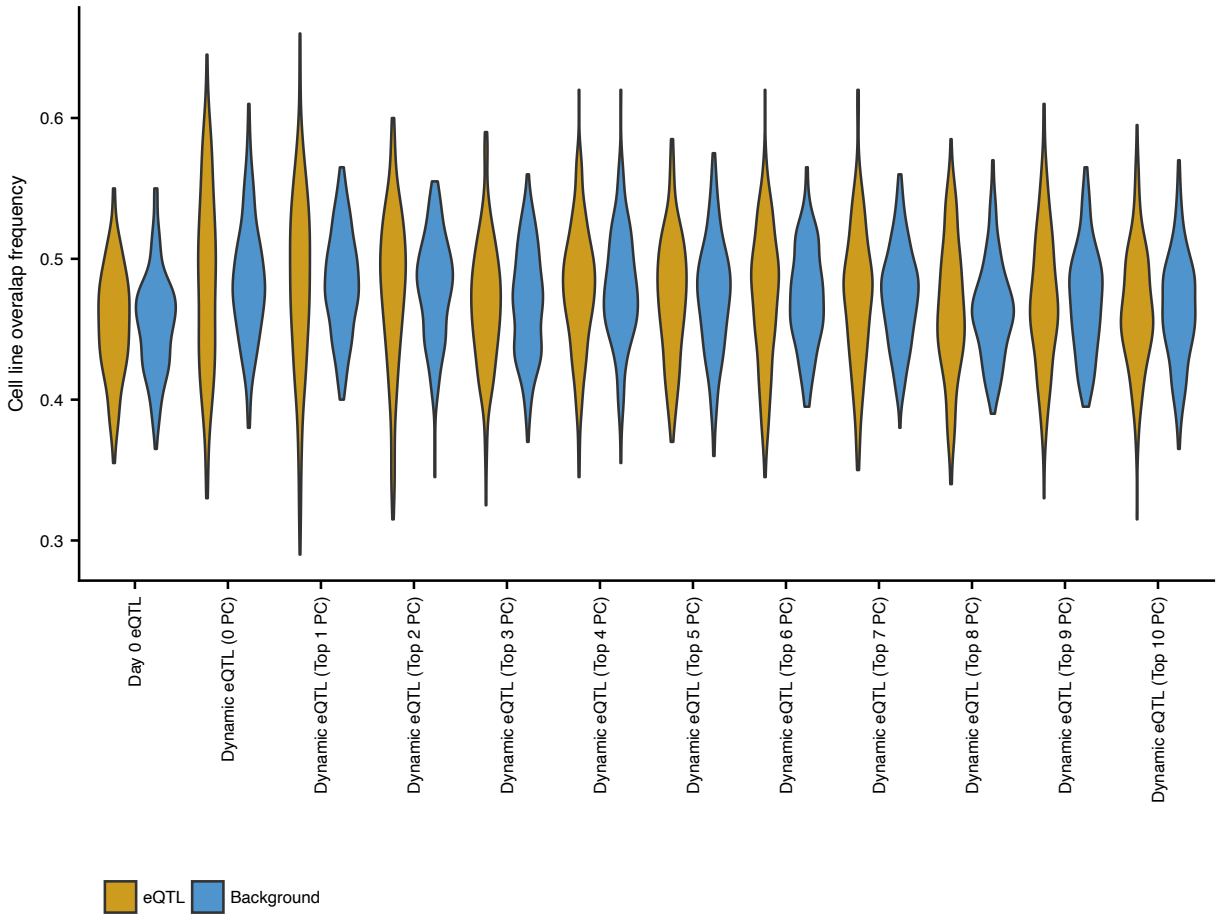


Fig. S2-16. Frequency of cell line overlap in genotype bins. Frequency at which each cell line pair is in the same genotype bin ({0,1,2}) across the strongest associated variants of the 200 most significant eQTL genes (gold) compared to MAF-matched randomly selected background variants (blue). Analysis 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.

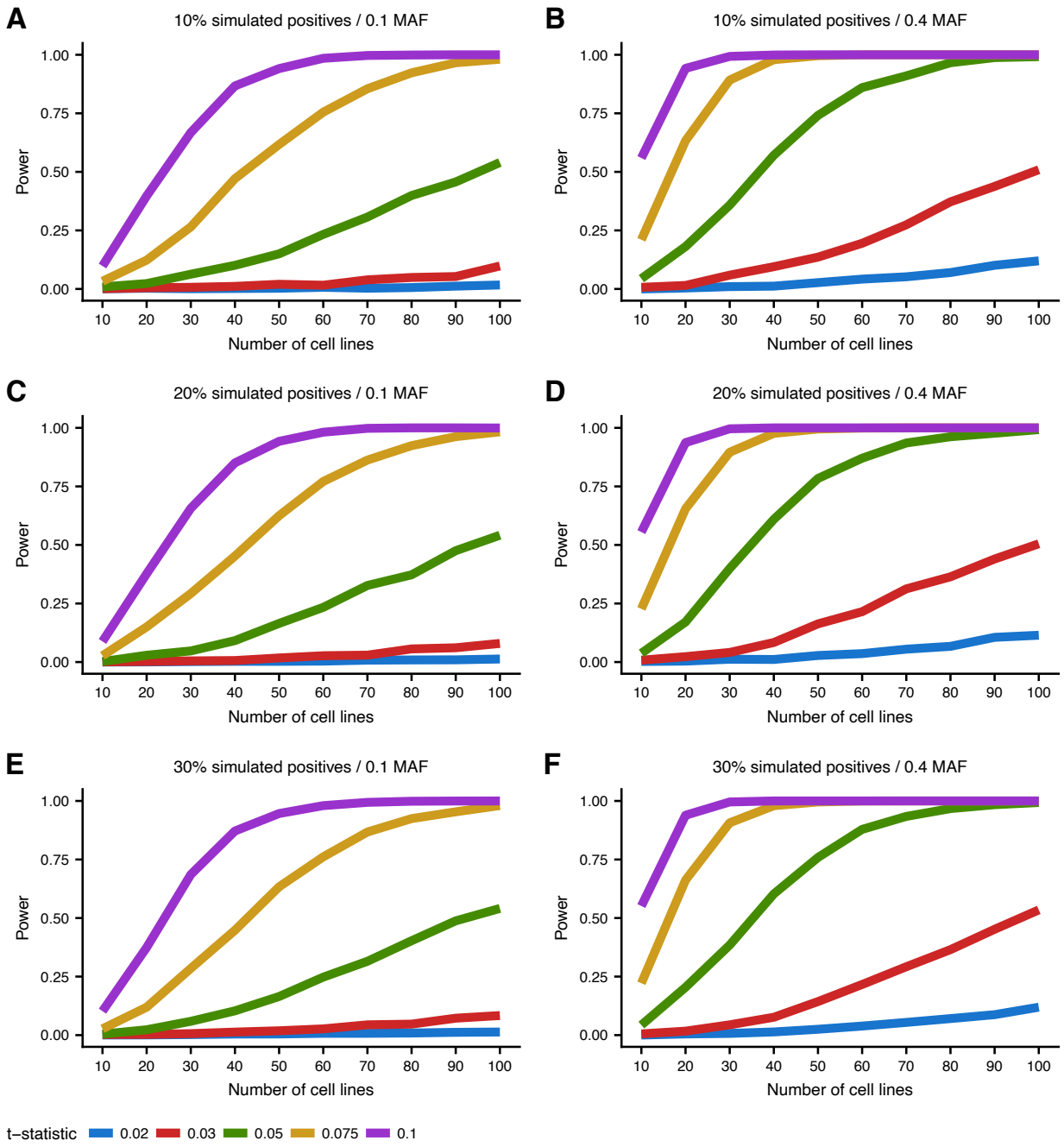


Fig. S2-17. Simulated power analysis for linear dynamic eQTLs. Power to detect simulated linear dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and 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).

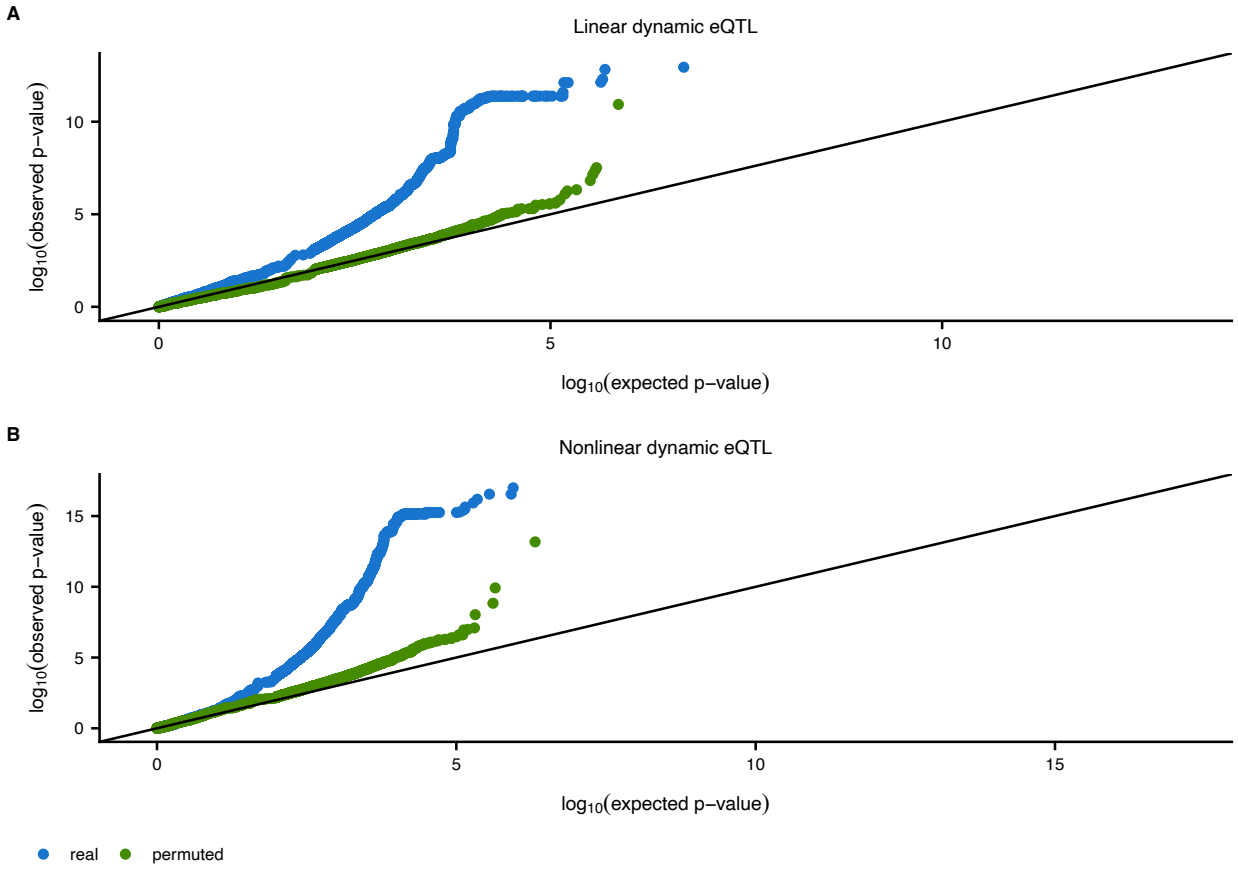


Fig. S2-18. Q-Q plots for linear and non-linear dynamic eQTLs. Q-Q plot for (A) linear and (B) non-linear dynamic eQTLs. Blue dots correspond to p-values from actual data relative to uniformly distributed p-values, whereas green dots correspond to p-values from permuted data relative to uniformly distributed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

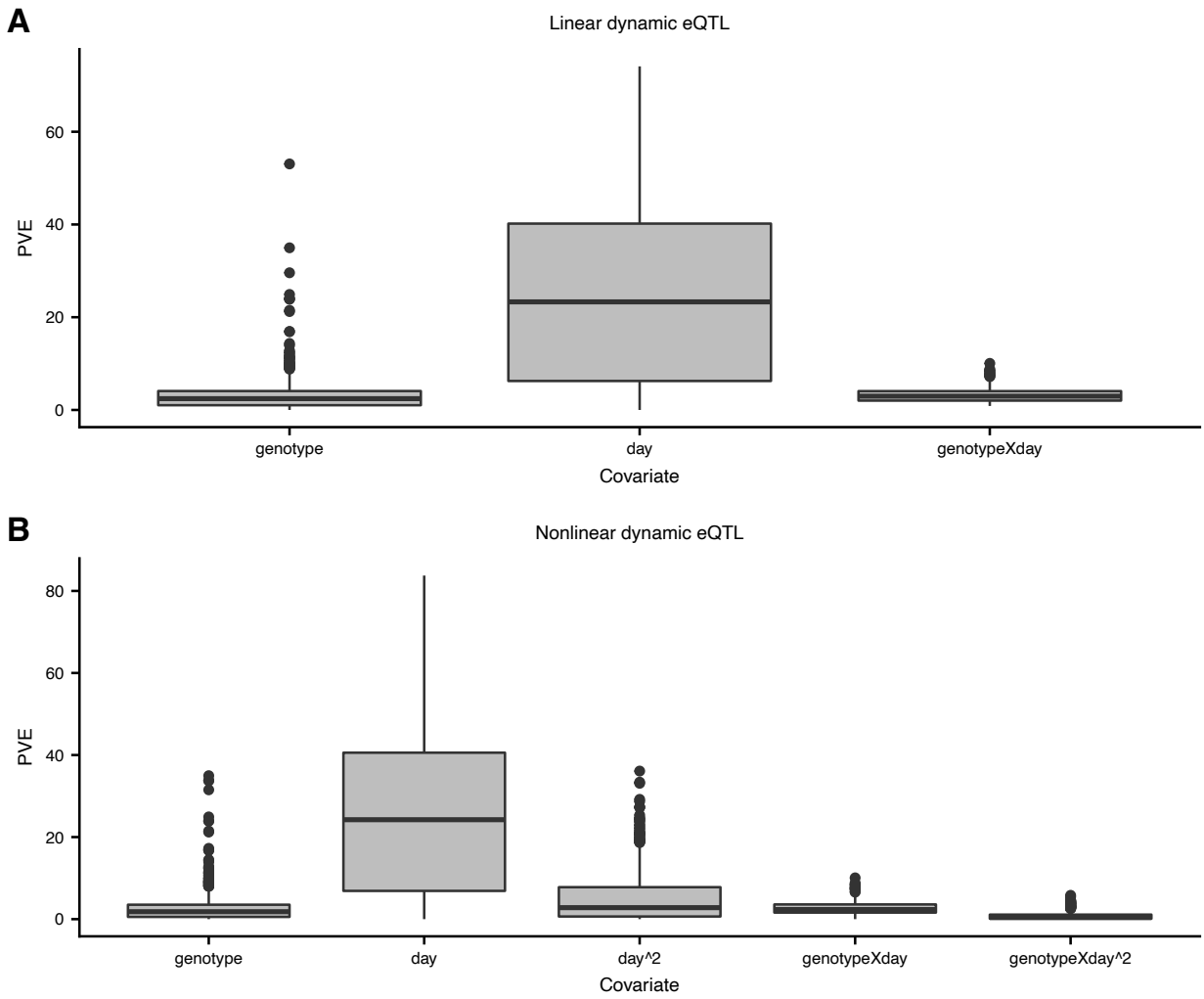


Fig. S2-19. Percent variance explained of dynamic eQTL covariates. Distribution of percent variance explained (PVE; y-axis) of each covariate (x-axis) across significant (eFDR <= .05) (A) linear dynamic eQTLs and (B) nonlinear dynamic eQTLs. For linear dynamic eQTLs, the interaction term (genotypeXday) 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.

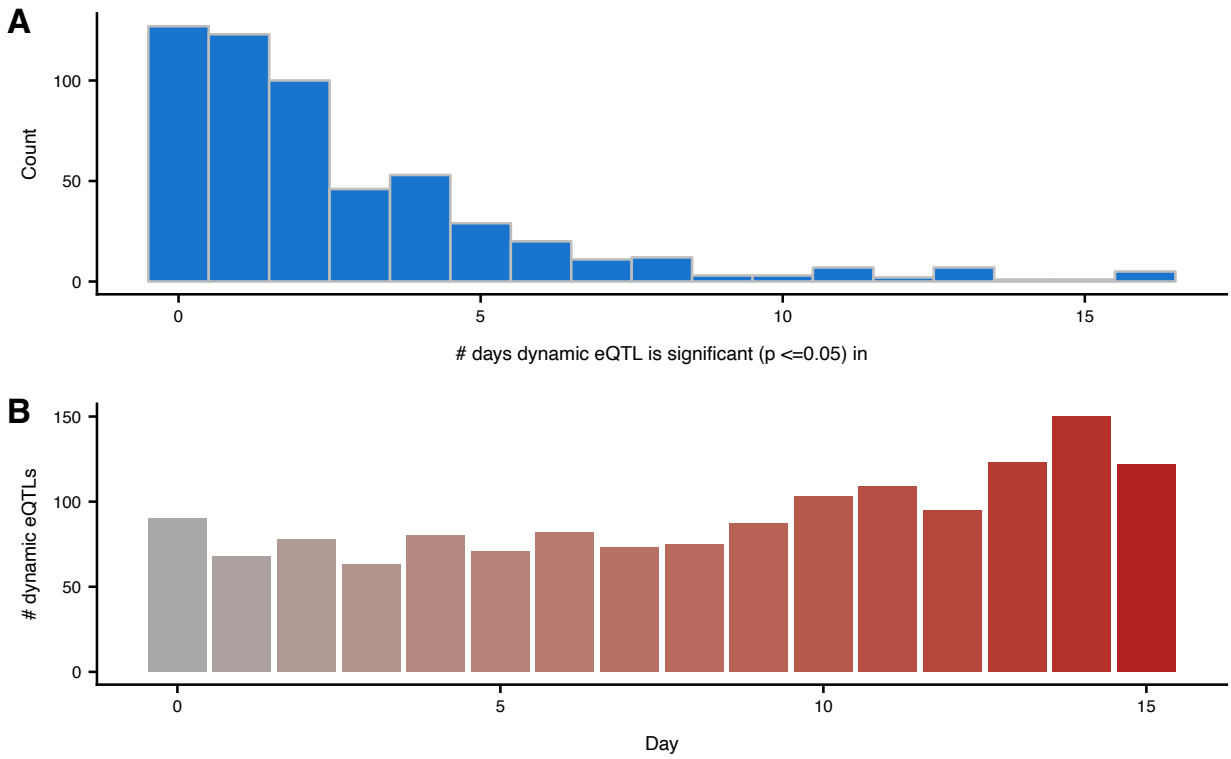


Fig. S2-20. Comparing linear dynamic eQTLs to non-dynamic eQTLs. (A) The number of time points in which the dynamic eQTLs (most significant variant per dynamic eQTL gene) have a nominally significant (p <= .05) non-dynamic eQTL. (B) The number of dynamic eQTLs (most significant variant per dynamic eQTL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

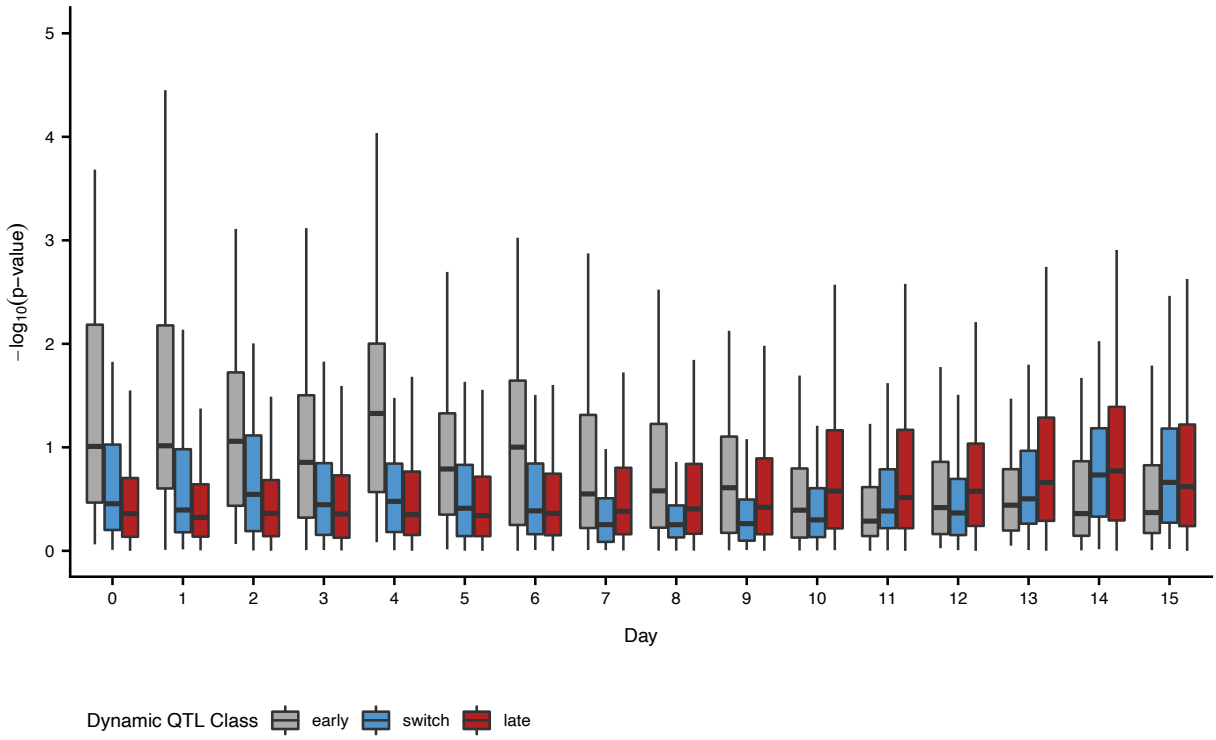


Fig. S2-21. Comparing linear dynamic eQTLs with non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of linear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by linear dynamic eQTL classifications (early, switch, and late).

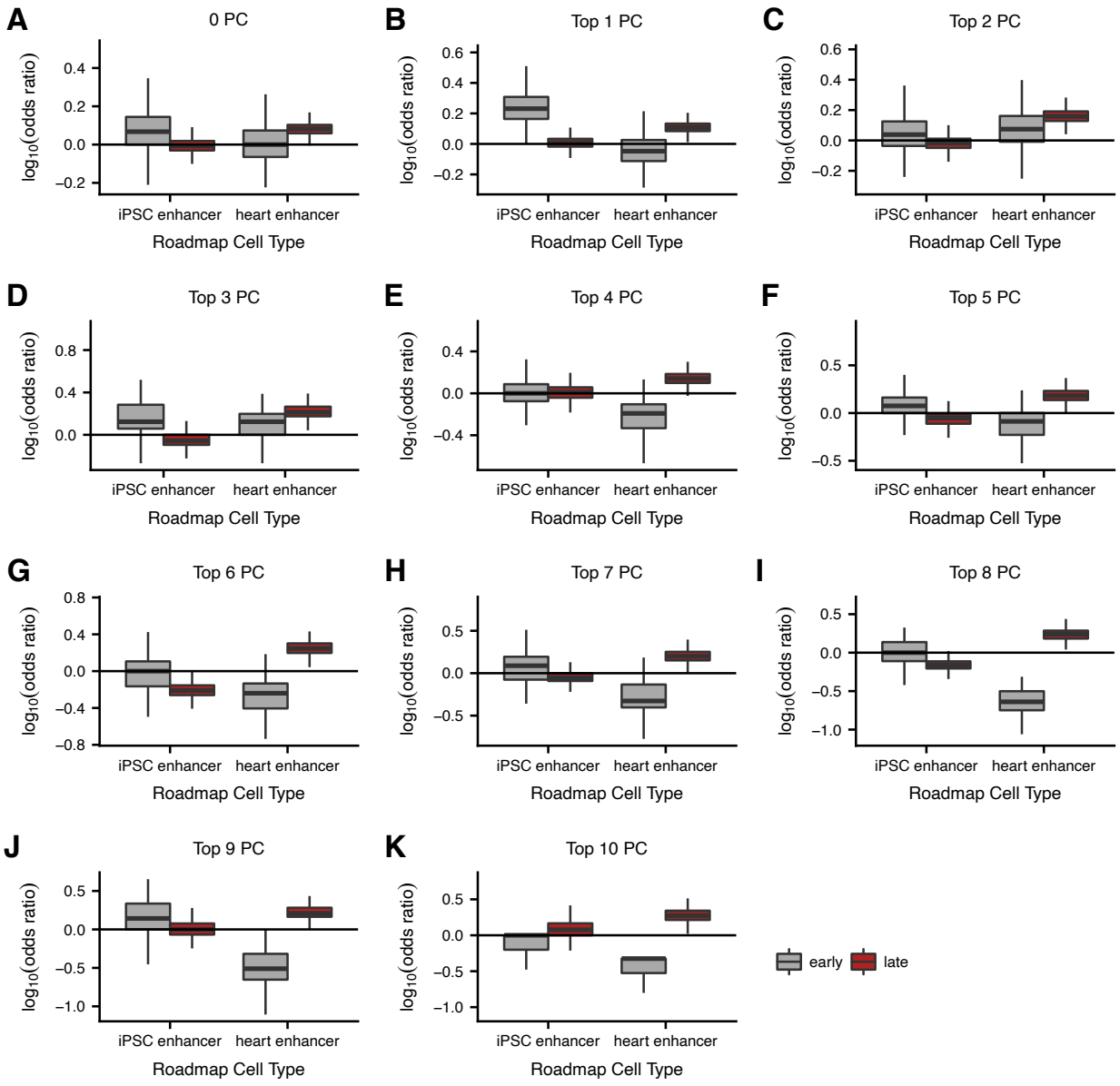


Fig. S2-22. Dynamic eQTL enhancer enrichment. Enrichment of dynamic eQTLs within cell type specific chromHMM enhancer elements relative to 1000 sets of randomly selected background variants matched for distance to transcription start site and minor allele frequency. Dynamic eQTLs were classified as early (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).

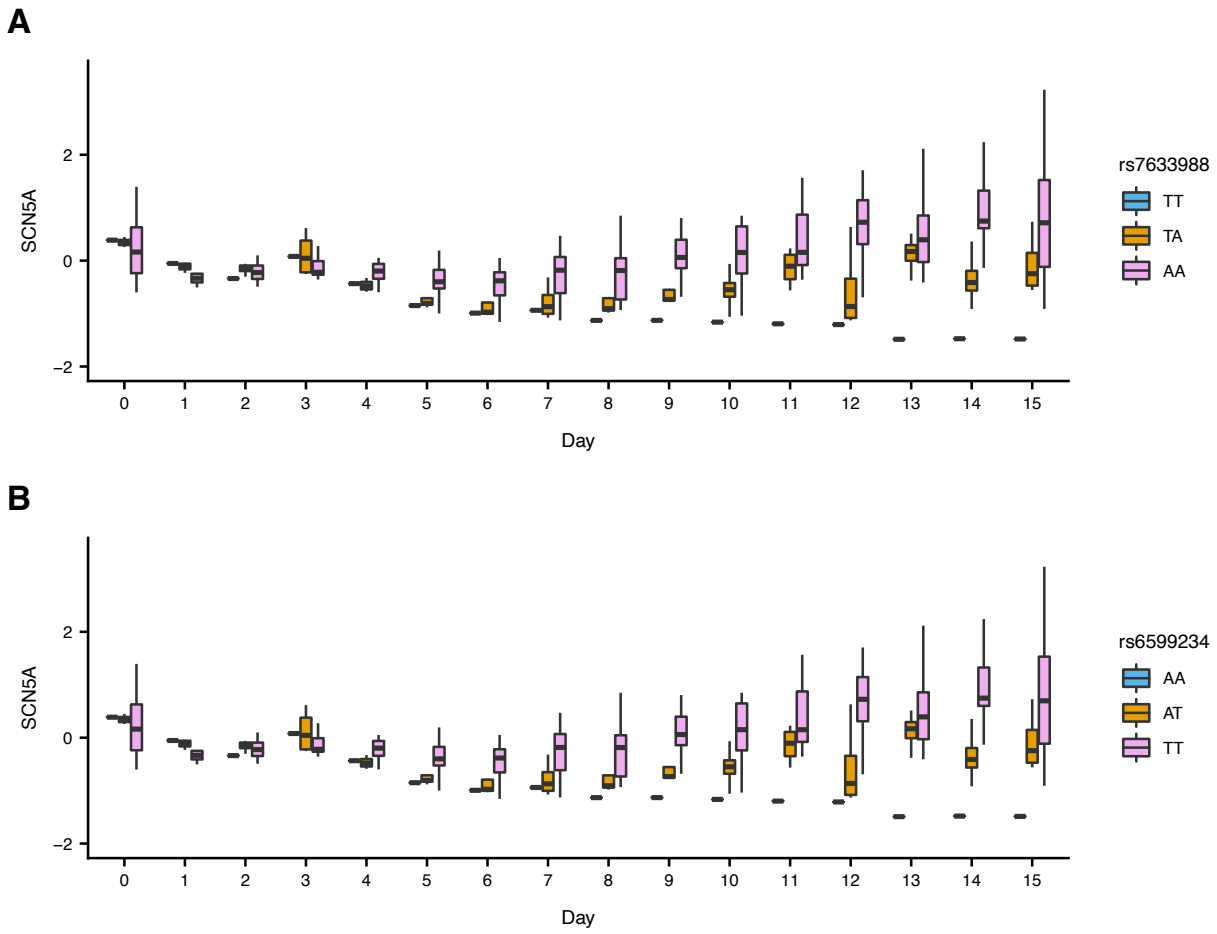


Fig. S2-23. Two significant linear dynamic eQTLs are known GWAS variants. Linear interaction association between time point (x-axis) and genotype (color) of (A) rs7633988 and (B) rs6599234 on residual gene expression (cell line effects regressed on expression) of *SCN5A* (y-axis).

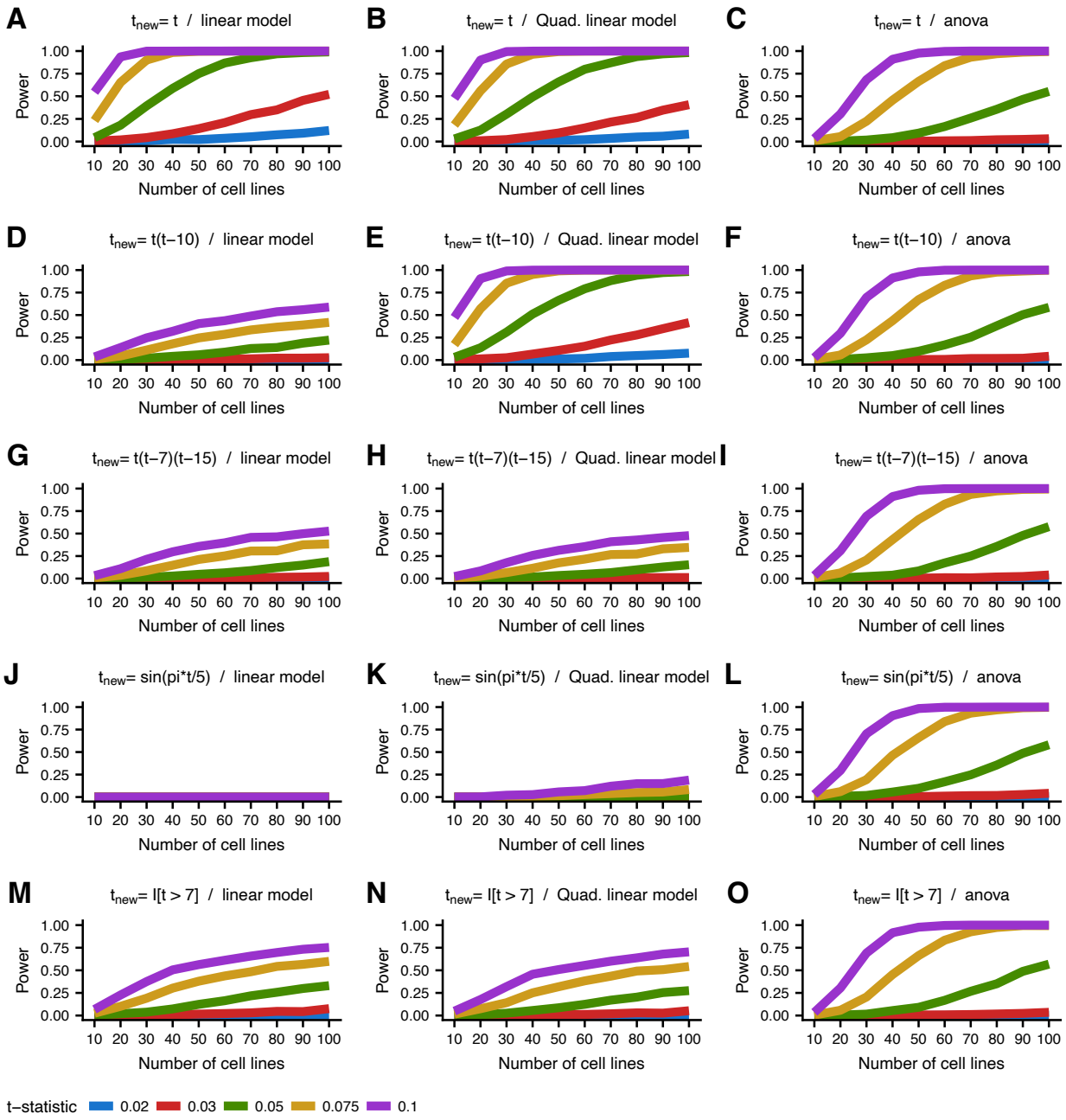


Fig. S2-24. Non-linear simulated power analysis. Power to detect simulated dynamic eQTLs (y-axis) based on 10,000 simulations at p-value <= 0.00017 (threshold corresponding to eFDR <= .05 for linear dynamic eQTLs in actual data) as a function of number of cell lines (x-axis) and t-statistic (color). 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.

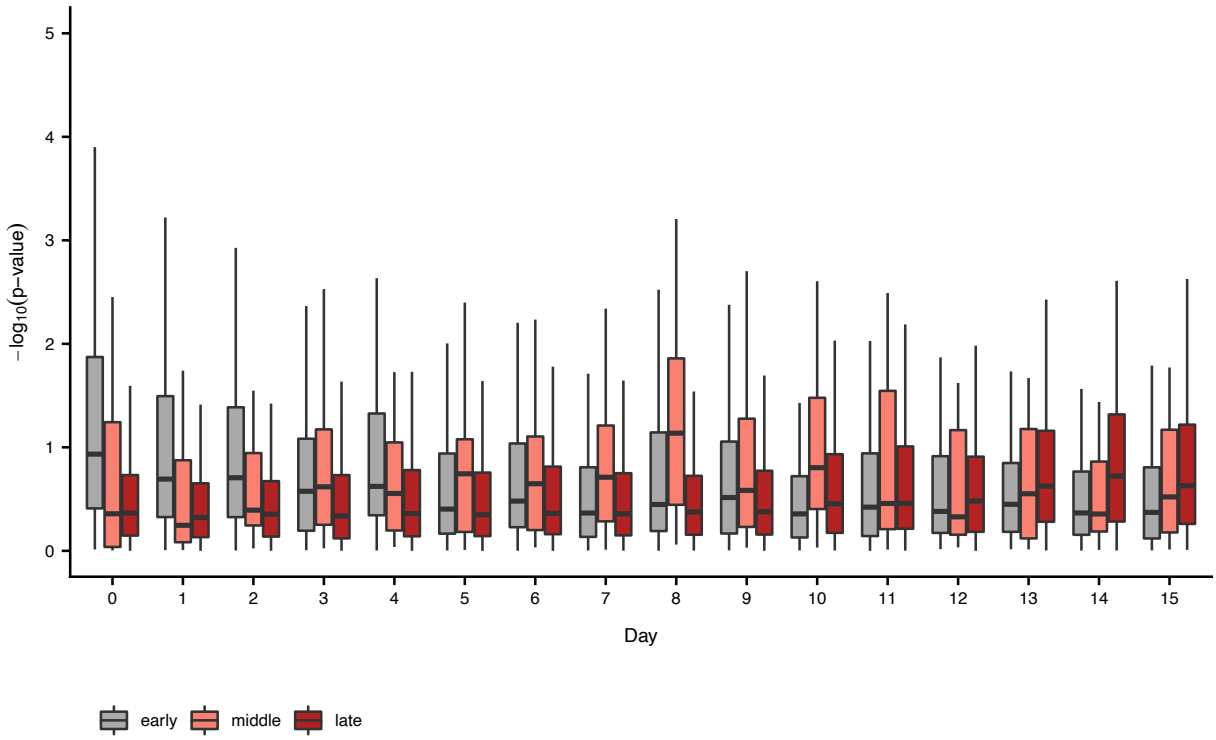


Fig. S2-25. Comparing nonlinear dynamic eQTLs to non-dynamic eQTLs. Non-dynamic eQTL p-values (y-axis) in all 16 time points (x-axis) of nonlinear dynamic eQTLs (most significant variant per dynamic eQTL gene) stratified by nonlinear dynamic eQTL classifications (early, middle, and late).

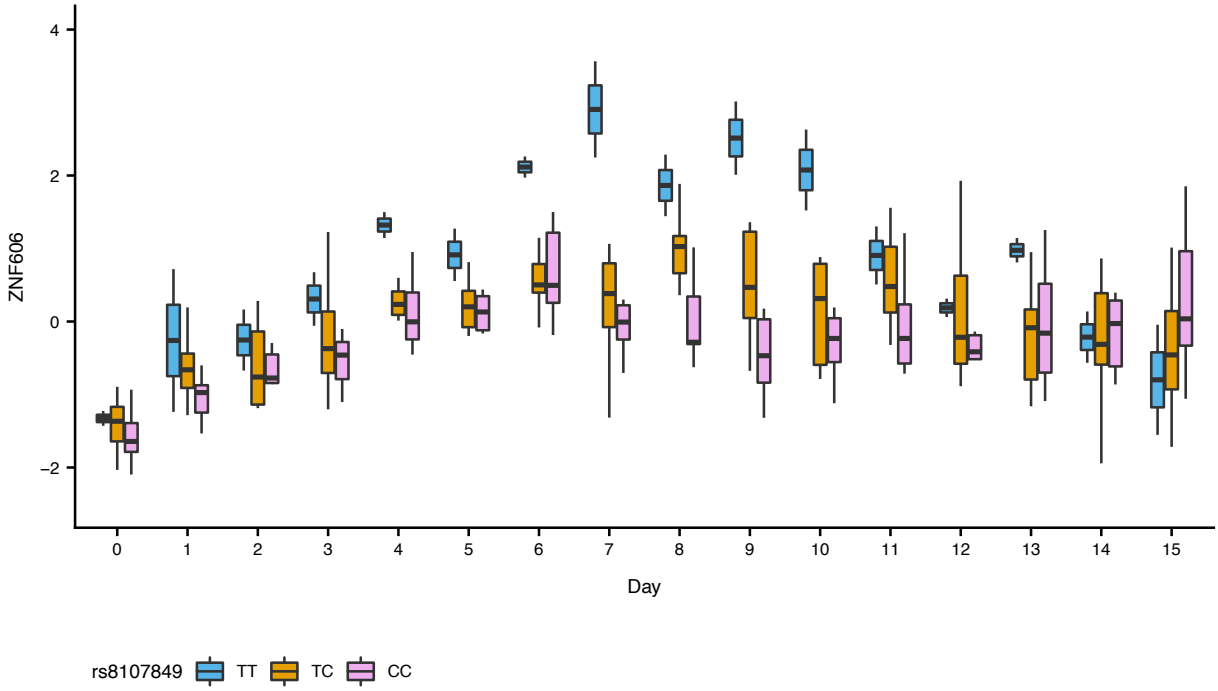


Fig. S2-26. Middle dynamic eQTL example. Nonlinear interaction association between genotype (color) of rs8107849 and time point (x-axis) on residual gene expression (cell line effects regressed on expression) of *ZNF606* (y-axis).

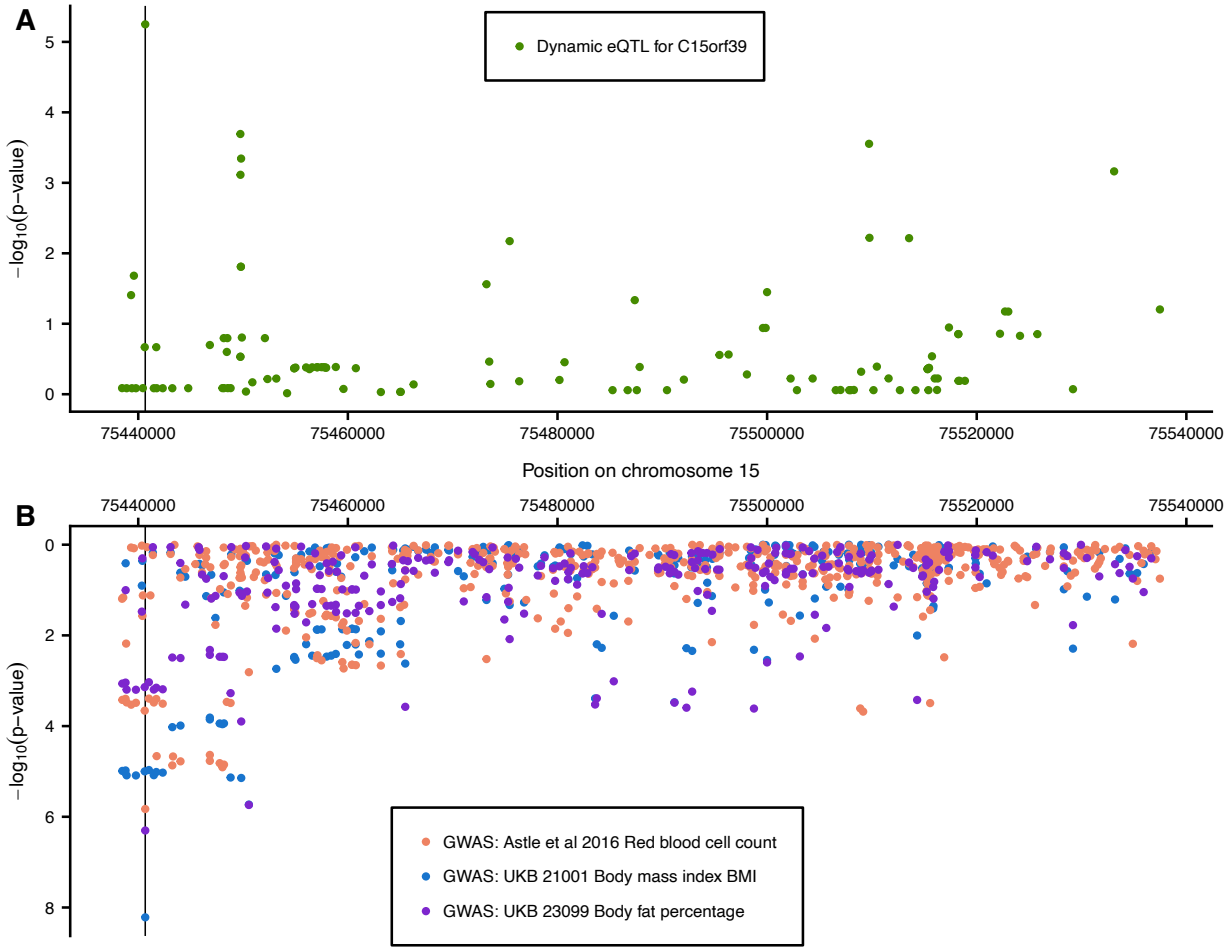


Fig. S2-27. Nonlinear dynamic eQTL overlaps GWAS variant. (A) Manhattan plot showing interaction association p-values for *C15orf39* according to nonlinear dynamic eQTL calling for all variants tested within 50KB of the *C15orf39* transcription start site. (B) Manhattan plot showing GWAS p-values on 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.

Supplementary Figures for Chapter III

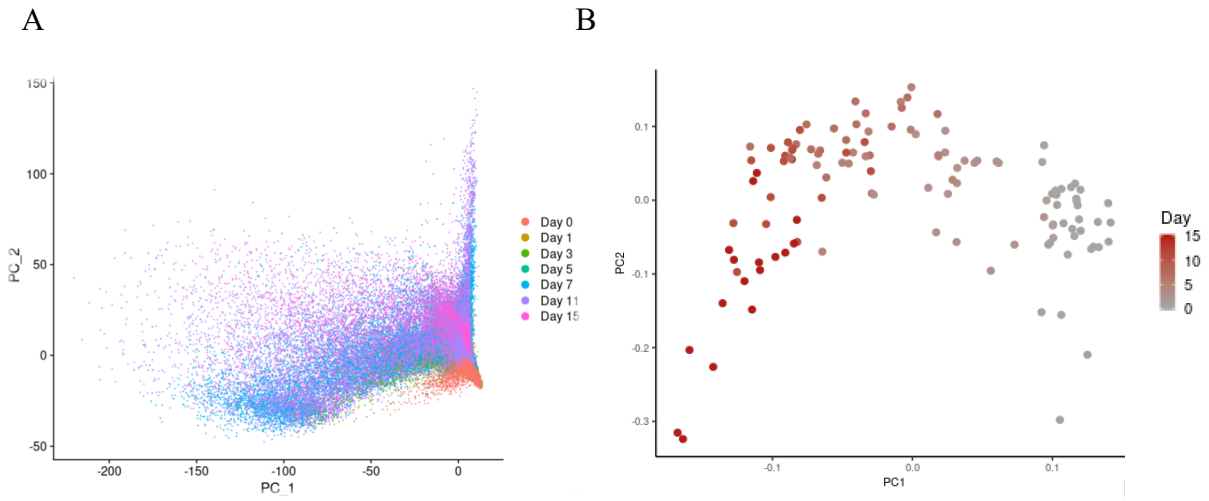


Fig. S3-1. Principal component analysis of single cell data. (A) PCA on full single cell dataset; cells are colored by differentiation day. (B) PCA on single cell data aggregated into pseudobulk by sample, where a sample refers to a given cell line at a given differentiation day. Samples colored on a gradient by differentiation day.

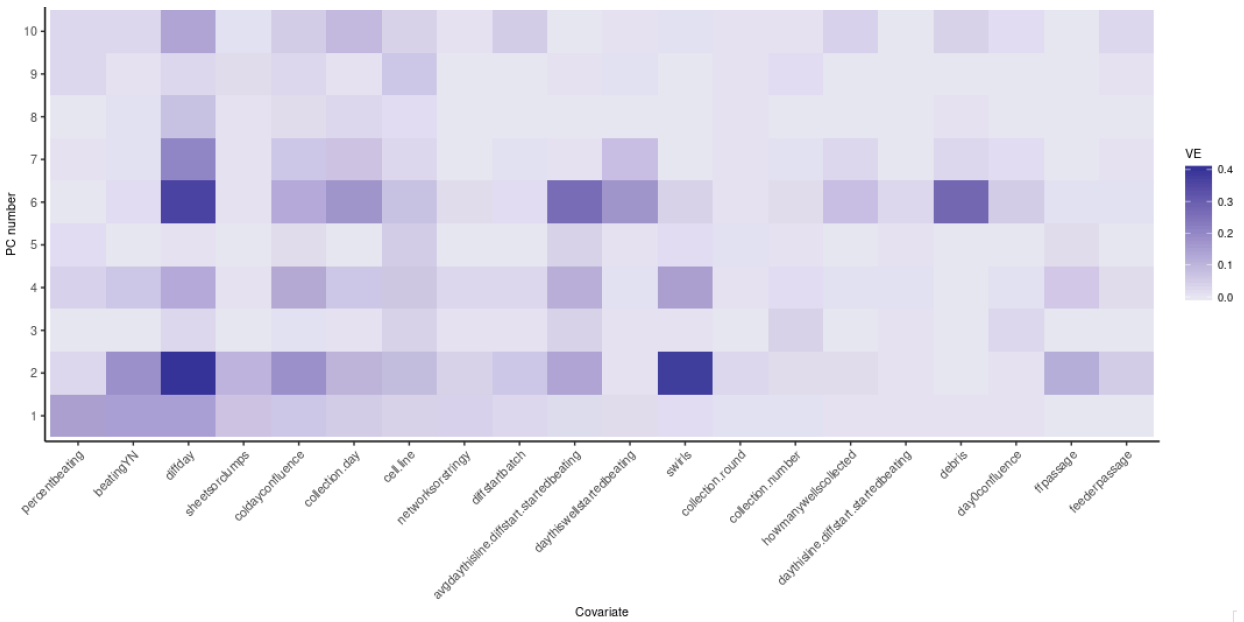
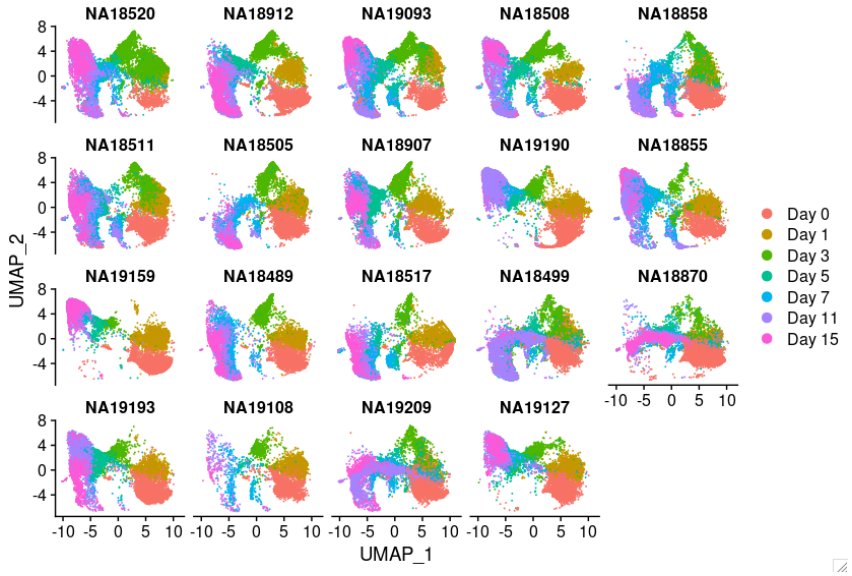


Fig. S3-2. Percent variance explained by technical factors in single cell data. Variance explained of each gene expression principal component (1-10) using recorded covariates, including: percent cells beating (visually assessed), differentiation day, collection day, culture confluence, cell morphology (visually assessed), and cellular debris.

A



B

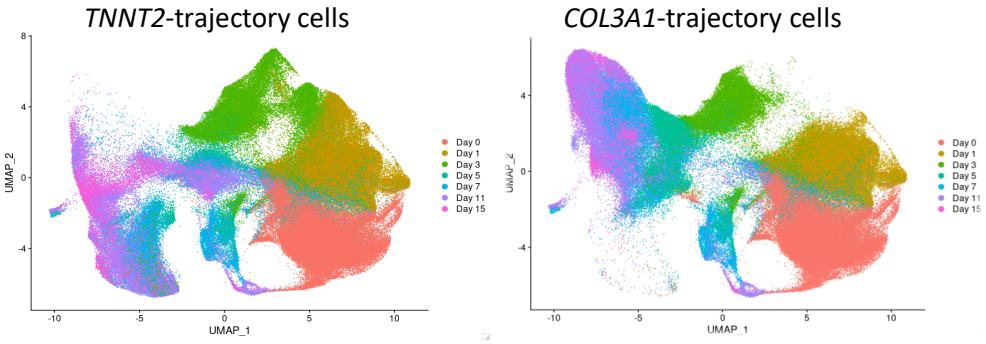


Fig. S3-3. Distinct cell trajectory groups in single cell data. (A) Each panel shows a UMAP of all cells for a given individual cell line. Cells are colored by differentiation day. (B) UMAP displaying cells of all individuals, where cells are separated into *TNNT2*-trajectory group (left) and *COL3A1*-trajectory group (right). Cells are colored by differentiation day.

Supplementary Tables for Chapter III

Col	Cell Line	Differentiation Day															
1	A	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	C					0	1	2	3	4	5	6	7	8	9	10	11
	D							0	1	2	3	4	5				
	B																0
2	B	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	D					0	1	2	3	4	5	6	7	8	9	10	11
	E							0	1	2	3	4	5				
	C																0
3	C	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	E					0	1	2	3	4	5	6	7	8	9	10	11
	F							0	1	2	3	4	5				
	D																0
4	D	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	F					0	1	2	3	4	5	6	7	8	9	10	11
	A							0	1	2	3	4	5				
	E																0
5	E	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	A					0	1	2	3	4	5	6	7	8	9	10	11
	B							0	1	2	3	4	5				
	F																0
6	F	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	B					0	1	2	3	4	5	6	7	8	9	10	11
	C							0	1	2	3	4	5				
	A																0

Table S3-1. Differentiation and Drop-seq batch collection schedule. Sample study design for a single differentiation and collection batch of 6 individual cell lines. Each line begins three staggered differentiations at three different starting days ("Day 0"). Samples are collected during three dedicated collection days (orange), where each collection contains three different cell lines at three differentiation days.