Single-Cell Sequencing Reveals Lineage-Specific Dynamic Genetic Regulation of Gene Expression During Human Cardiomyocyte Differentiation

Authors: Reem Elorbany[1*], Joshua M Popp[2*], Katherine Rhodes[3], Benjamin J Strober[2], Kenneth Barr[3], Guanghao Qi[2], Yoav Gilad[3,4**], Alexis Battle[2,5**]

Affiliations: 1. Interdisciplinary Scientist Training Program, University of Chicago, Chicago, IL 60637, USA. 2. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. 3. Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. 4. Department of Medicine, University of Chicago, Chicago, IL 60637, USA. 5. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

* These authors contributed equally to this work

** **Co-corresponding authors**

Abstract

Dynamic and temporally specific gene regulatory changes may underlie unexplained genetic associations with complex disease. During a dynamic process such as cellular differentiation, the overall cell type composition of a tissue (or an *in vitro* culture) and the gene regulatory profile of each cell can both experience significant changes over time. To identify these dynamic effects in high resolution, we collected single-cell RNA-sequencing data over a differentiation time course from induced pluripotent stem cells to cardiomyocytes, sampled at 7 unique time points in 19 human cell lines. We employed a flexible approach to map dynamic eQTLs whose effects vary significantly over the course of bifurcating differentiation trajectories, including many whose effects are specific to one of these two lineages. Our study design allowed us to distinguish true dynamic eQTLs affecting a specific cell lineage from expression changes driven by potentially non-genetic differences between cell lines such as cell composition. Additionally, we used the cell type profiles learned from single-cell data to deconvolve and re-analyze data from matched bulk RNA-seq samples. Using this approach, we were able to identify a large number of novel dynamic eQTLs in single cell data while also attributing dynamic effects in bulk to a particular lineage. Overall, we found that using single cell data to uncover dynamic eQTLs can provide new insight into the gene regulatory changes that occur among heterogeneous cell types during cardiomyocyte differentiation.

Introduction

A primary aim of human genetics and genomics is to understand the genetic architecture of complex traits. Current studies demonstrate that the majority of trait-associated genomic loci are in non-coding regions of the genome, and are thought to be involved in gene regulation (Edwards et al. 2013). Therefore, studies exploring gene regulation are essential to our understanding of complex phenotypes (Li et al. 2016, Albert et al. 2015). Studies mapping expression quantitative trait loci (**eQTLs**), identifying genetic variants associated with gene expression levels, reveal the impact of genetic variation on gene regulation and can inform molecular mechanisms underlying trait-associated loci. eQTLs have now been identified for a wide variety of tissues, and their study has contributed to the understanding of gene regulation and disease (GTEx Consortium 2020; Lappalainen et al. 2013; Battle et al. 2014; Pickrell et al. 2010; Stranger et al. 2012; Nica et al. 2010; Nicolae et al. 2010).

Gene regulation, including genetic regulation of gene expression, can vary between contexts including different cell types, temporal stages, and environmental stressors. Particular attention has been paid to differences in gene regulation between tissues and cell types. Large studies including the Genotype-Tissue Expression Project (**GTEx**) have been now been successful in identifying thousands of eQTLs in diverse human tissues [GTEx Consortium 2020; Nica et al. 2011). However, despite these efforts, we are still unable to identify a regulatory mechanism for the genetic contribution of a majority of disease-associated loci (Bis et al. 2011, Myocardial Infarction Genetics Consortium 2009, Manolio et al. 2009, Eichler et al. 2010, Arvanitis et al. 2020). One reason for this knowledge gap may be that most large-scale eQTL studies are based on expression data from adult, bulk tissue samples that do not represent the specific cell types and contexts in which disease-relevant dysregulation occurs (Umans 2020).

Recent advances in single-cell sequencing have allowed us to assay gene expression in individual cells, allowing us to access disease relevant cell types and cell states, even if they compose a small fraction of a tissue and would not be well captured by bulk data, and even if they are not known a priori. Indeed, single cell datasets have revealed a more complex landscape of gene expression in individual cell types than previously known in tissues such as brain and kidney (Welch et al. 2019, Park et al. 2018). Likewise, mapping eQTLs from single-cell RNA-sequencing data promises to enable the identification of previously undiscovered disease-relevant regulatory mechanisms. Recently, collection and analysis of population-scale scRNA-seq datasets have demonstrated that genetic effects do vary between cell types belonging to the same tissue (Fairfax et al. 2012, Kasela et al. 2017, Kim-Hellmuth et al. 2020).

Beyond cell-type specificity, only a small number of studies have attempted to characterize dynamic gene regulatory changes that occur during development or among contexts that change over time (Strober et al 2019, Knowles et al 2017, Taylor et al 2018, Fairfax et al 2014, Smirnov et al. 2009; Watts et al. 2002, Kariuki et al. 2016; Alleyne et al. 2017). These have highlighted temporally specific eQTL effects that were not evident from static data. Studying the temporal dynamics of gene expression has the potential to uncover genomic loci involved in gene regulation during developmental processes and identify associations that were previously overlooked. Accordingly, we previously studied genetic effects on the regulation of gene expression during the differentiation of induced pluripotent stem cells (**iPSCs**) to cardiomyocytes (Strober et al 2019). We collected time-series bulk RNA-seq data for nineteen individuals to identify hundreds of eQTLs displaying dynamic, and sometimes transient effects on expression across the course of cardiomyocyte differentiation. These dynamic eQTLs included genetic variants which were associated with cardiovascular disease-related traits, including obesity.

However, the complexities of cardiomyocyte differentiation and other dynamic processes are not fully captured by bulk RNA-seq data even in a time course study design. During development and differentiation, expression profiles change over time in individual cells along a spectrum of maturity (Pijuan-Sala et al 2018). Cells within a single sample do not necessarily differentiate at the same rate, along the same trajectory, or even toward the same terminal cell type. Different cell lines may also vary in the proportion of cells in different states at each time point. Indeed, in our previous work, we identified two clusters of cell lines undergoing cardiomyocyte differentiation that exhibited broad differences in the expression trajectory of groups of genes over time (Strober et al 2019). Bulk expression profiles represent an average across cells from various points across a developmental landscape, obscuring the underlying variation in cell state, and even making it difficult to definitively attribute differences to cis-regulatory genetic effects. Recent work has demonstrated that the improved resolution of single-cell RNA-seq data can identify homogeneous subpopulations of cells at similar stages of differentiation, offering a clearer view of genetic regulation in an individual time step (Cuomo et al. 2020, Jerber et al. 2021). However, such analysis has only been applied to a few cell types, not including cardiomyocytes, and has been limited to the study of dynamics within a single lineage.

In this study, we applied single-cell RNA-seq to the nineteen cell lines assayed in our previous bulk RNA-seq analysis, collecting single-cell data at seven informative time points during

cardiomyocyte differentiation, enabling us to observe cell-type specificity, cell composition differences, and temporal changes together in a unified experiment. The resolution of this single cell data enables us to characterize the cardiomyocyte differentiation landscape in much greater detail than was possible in bulk. We identify a bifurcation in cell fate, which explains the previously observed clustering of cell lines and enables us to study genetic regulatory dynamics along two distinct trajectories with a single experiment. Characterization of these trajectories allows us to reanalyze existing bulk samples and mitigate confounding impact of cellular composition and identify dynamic effects specific to each lineage (Westra et al. 2015, Kim-Hellmuth et al. 2020).

Results

We differentiated induced pluripotent stem cells (**iPSCs**) from 19 human cell lines into cardiomyocytes; these same cell lines were previously used for a cardiomyocyte time course study published in Strober et al 2019. For the current study, we used new iPSC cultures of the same lines, and differentiated them again to cardiomyocytes. We used Drop-seq to collect single-cell RNA-seq data at 7 days throughout the 16-day differentiation time course. We chose to collect data from days 0 (iPSC), 1, 3, 5, 7, 11, and 15 (cardiomyocyte), as we have previously observed that these days represent the most informative stages during this particular differentiation trajectory (Strober et al. 2019, Selewa et al. 2020). We collected single-cell data using a balanced study design in which each collection included three individuals at three unique differentiation time points. This design minimizes technical effects associated with individual and differentiation day. After filtering data from low quality cells (Methods), the resulting 131 samples contained an average of 1,762 cells per sample and an average of 1,375 genes detected as expressed per cell. Following normalization, a principal component analysis revealed that, as expected, differentiation day is the primary axis of variation in the single cell gene expression data (**Fig. S1a-b**).

Differentiation progress and cell line differences drive variation in gene expression

In order to characterize the complex landscape of cardiomyocyte differentiation, we used UMAP to produce a low-dimensional embedding of the single cell data while preserving global structure. We found that while cells from the early days of the differentiation time course exhibited fairly uniform transcription profiles, this was less true for later days (days 7, 11, and 15; **Fig. 1A, 1D**). Marker genes known to be expressed at various stages in cardiac differentiation, from iPSC to mesoderm to cardiomyocyte, showed high expression at expected early, intermediate, and late stages of the differentiation time course, respectively (**Fig. 1B, 1E**). Next, we used unsupervised clustering to identify distinct cell populations present in the data, and matched these to known cell types based on expression of known marker genes (**Fig. 1C**, Methods, Burridge et al. 2014**)**. As suggested by previous reports (Strober et al. 2019, Selewa et al. 2020), we identified a bifurcation in the differentiation landscape, giving rise to two distinct terminal cell types. One of these terminal cell types 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, Fig. 1B). Cells in the other terminal cell type do not express cardiomyocyte markers, and instead have high expression of genes such as *COL3A1* and *VIM*, which are expressed in the extracellular matrix of cardiac fibroblasts (Ieda et al. 2009, Zhang et al. 2019). The differentiation outcome of each sample, namely the proportion of cells in each cluster, varied by individual cell line; certain lines differentiated primarily into either the *TNNT2*-expressing or the *COL3A1*-expressing terminal cell type clusters (**Fig. 1E, S2**). For the remainder of this paper, we will refer to the *TNNT2*-expressing cell cluster as cardiomyocytes (CM) and to the *COL3A1*-expressing cluster as cardiac-fibroblasts (CF) or fibroblast. We also identified a cluster that underexpressed marker genes of cardiac cell types throughout the differentiation process, and instead expressed several endoderm-specific markers such as *APOA1* and *AFP*. We were unable to fully characterize this cluster based on expression patterns alone, and omitted these cells from downstream investigation of the dynamics of gene regulation on gene expression during mesoderm and cardiac cellular differentiation.

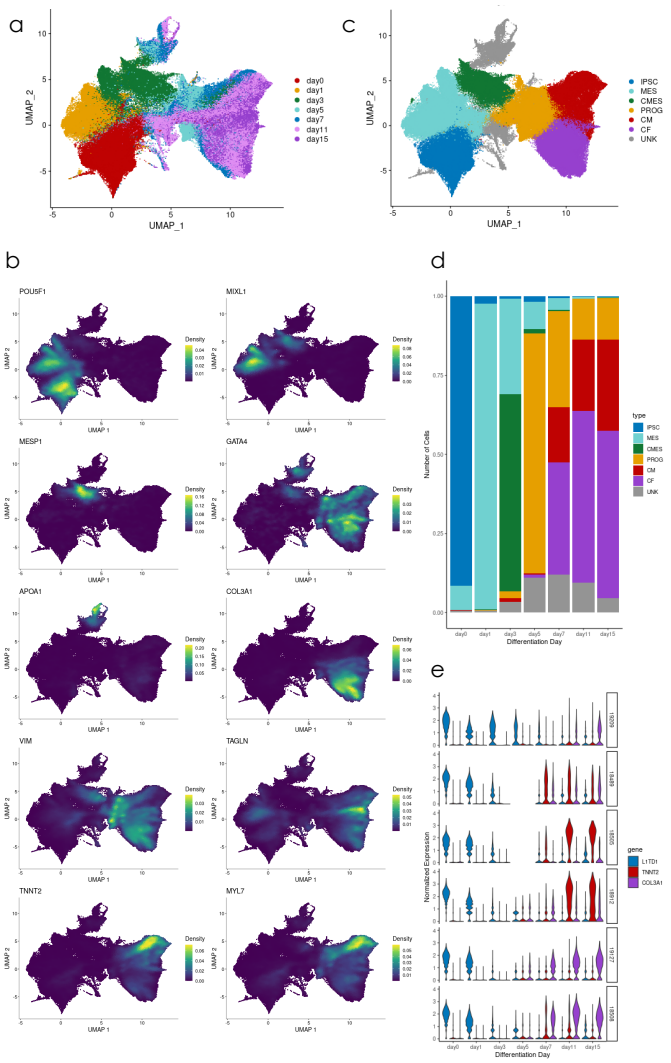


Fig. 1. Gene expression patterns in single cell data. (A) UMAP of full single cell dataset; cells are colored by differentiation day. (B) Estimated density of expression for several marker genes across cells. (C) UMAP of full single cell dataset; cells are colored by cell type, assigned based on Leiden clustering and marker gene expression. IPSC=induced pluripotent stem cell, MES=mesoderm, CMES=cardiac mesoderm, PROG=cardiac progenitor, CM=cardiomyocyte, CF=cardiac fibroblast, UNK=unknown cell type. (D) Proportion of cells belonging to each cell type per differentiation day, across all cell lines. (E) Distribution of *L1TD1* (pluripotency marker), *TNNT2* (cardiomyocyte marker) and *COL3A1* (cardiac fibroblast marker) over cells from 6 representative examples of the 19 cell lines studied, for each of the 7 differentiation days.

Single-cell data offers a highly resolved view of cellular differentiation

In previous work, we investigated the relationship between genotype and chronological time, represented by the differentiation day in which each bulk sample was collected. However, chronological time may not properly capture the axis of variation along which genetic regulation is changing, and can be heavily confounded by heterogeneity in differentiation within and between samples. If cells within a sample progress through differentiation at different rates, their aggregated expression profile will not be truly reflective of

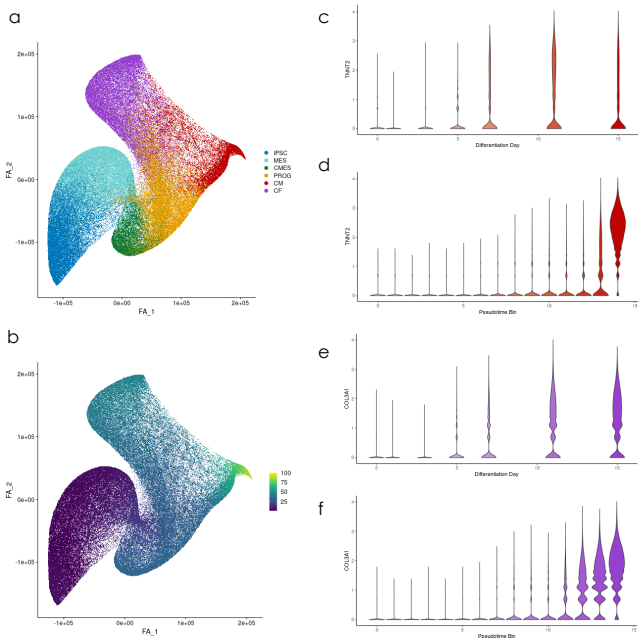


Fig. 2. Pseudotime inference and pseudobulk aggregation. (A) Force atlas embedding of all cells from the two cardiac differentiation lineages combined, colored by cell type. (B) Force atlas embedding from (A), colored by pseudotime, which was inferred for each cell shown using diffusion pseudotime. (C) Distribution of normalized expression of *TNNT2*, a cardiomyocyte marker gene, across cells from the cardiomyocyte lineage for each differentiation day. (D) Normalized *TNNT2* expression across cells from each of 16 pseudotime quantile bins along the cardiomyocyte trajectory. (E) Normalized expression of *COL3A1*, a cardiac fibroblast marker, across cells from the cardiac fibroblast lineage for each differentiation day. (F) *COL3A1* expression across cells for 16 pseudotime quantile bins along the cardiac fibroblast trajectory.

an individual stage of differentiation, confounding tests for association between genotype and differentiation progress. Systematic differences between cell lines can exaggerate this: differentiation speed appears to vary between cell lines (Fig. 1E), such that differentiation progress at day 3, for example, is not uniform across samples. Such differences can lead to false associations between genotype and differentiation progress in cases where genotype is partially correlated with a cell line's differentiation speed.

Cellular heterogeneity drives further challenges when aggregating across cells that are differentiating along diverging paths. Aggregated bulk profiles will lose information about the individual cell types present, and if cell type composition varies between individuals (Fig. 1E), this will further confound associations between genotype and expression changes during differentiation.

By collecting expression at the single-cell level, we are able to address both of these challenges. To properly focus on the two primary cardiac lineages present, we used the *scanpy* package to produce a low-dimensional Force Atlas embedding of the cells that had been successfully assigned to a known cell type (**Fig. 2A,** Wolf et al. 2018, Jacomy et al. 2014). We inferred pseudotime for each cell with diffusion pseudotime (Haghverdi et al. 2016, Wolf et al. 2019), so that progress through differentiation is learned from cells' individual expression profiles rather than their time of collection (**Fig. 2B**). With each cell assigned to a cell type (Fig. 1C), we are additionally able to account for diverging paths by studying gene regulatory dynamics within each lineage separately.

One disadvantage to single-cell data compared to bulk is that single-cell measurements are more sparse and noisy: by aggregating over cells, bulk RNA-sequencing reduces noise, which makes expression measurements more tractable for eQTL calling. We therefore partitioned cells (separately for each lineage) into pseudotime bins, pooling information across cells to mitigate the noisiness of single cell expression measurement while maintaining homogeneous populations of cells through lineage subsetting and pseudotime binning. This aggregation scheme enables us to

produce a greater number of samples, as we are no longer constrained to the 7 days when experimental collection was performed, while maintaining the expected trends of lineage-specific marker gene expression over pseudotime (**Figs. 2C-F**).

Mapping of dynamic eQTLs

We applied a Gaussian linear model to the aggregated single-cell pseudo-bulk data based on pseudotime bins from each lineage to identify dynamic eQTLs, namely variant-gene pairs in which the interaction effect of genotype and differentiation time is significantly associated with changes in gene expression. We identified linear dynamic eQTLs for 357 genes in the cardiomyocyte lineage ($q < 0.05$) and 903 genes in the cardiac fibroblast lineage (Methods; **Table 1**).

We found that both lineage specificity and the replacement of real chronological time with pseudotime improved power for dynamic eQTL detection. For comparison, using chronological differentiation day as the time variable identified only 142 and 29 dynamic eQTLs for the cardiomyocyte and cardiac fibroblast lineages, respectively. Using differentiation day as the time variable and omitting lineage specificity altogether identified only 5 dynamic eQTLs in the pseudobulk data. Ultimately, our lineage subsetting and pseudotime approach revealed more dynamic eQTLs than were previously identified in an experiment with bulk collections at over twice as many time points (Strober et al 2019). To ensure a meaningful comparison, we re-processed the previously collected bulk data in a similar pipeline as pseudo-bulk, accounting for changes in hypothesis testing and filtering of variant-gene pairs (Methods). This revealed a total of 1028 genes with a dynamic eQTL (compared to a total of 1056 genes detected between both lineages with pseudobulk binned to a similar number of samples). The increased detection rate may stem from increased homogeneity of cellular populations that undergo pseudo-bulk aggregation, as well as improved measurement of differentiation progress achieved by using cellular pseudotime rather than sample collection time.

As an example of the trait relevance of these dynamic eQTLs, one dynamic eQTL variant, rs1234988, has previously been implicated by GWAS to be associated with hypertension (p=2.5e-

Dataset	*Aggregation*	*Time Points*	*Lineage*	*Dynamic eGenes Detected*	*Total # Genes Tested*	*Total # Tests*
Pseudobulk	Pseudotime	16	CM	357	8,969	1,601,727
Pseudobulk	Pseudotime	16	CF	903	9,140	1,633,408
Pseudobulk	Differentiation Day	7	CM	142	9,541	1,693,532
Pseudobulk	Differentiation Day	7	CF	100	9,548	1,711,693
Pseudobulk	Differentiation Day	7	Combined	5	9,656	1,731,798
Bulk	Differentiation Day	7	Combined	210	10,772	1,963,378
Bulk	Differentiation Day	16	Combined	1028	10,981	1,991,072

Table 1. Comparison of dynamic eQTL calling methods. We report the number of dynamic eGenes (genes with a significant dynamic eQTL at gene-level q-value <= 0.05), for each of the aggregation schemes assessed. Total number of genes tested and total number of tests run are also reported.

35), and was detected as a dynamic eQTL for *ARHGAP42*, a Rho GTPase which has previously been identified as a critical regulator of vascular tone and hypertension in mice (**Fig. 3A-B**, Barbeira et al. 2021, Loirand and Pacaud 2014). Notably, *ARHGAP42* is known to be a smooth-muscle selective Rho GAP, and this dynamic eQTL was exclusively identified in the cardiac fibroblast lineage (Bonferroni-adj. p=2.4e-5, cardiac fibroblast lineage, adj. p=0.79, cardiomyocyte lineage). This variant is not detected as a dynamic eQTL without lineage subsetting or pseudotime binning (adj. p=1). This example illustrates the advantages of incorporating exploratory data analysis in the study of *in vitro* experimental datasets: while the differentiation procedure used for these experiments was designed to produce exclusively cardiomyocytes, an alternative terminal cell type discovered after exploratory data analysis is able to provide meaningful insight into an additional differentiation process.

The pseudotime values can be interpreted as intermediate time points with greater resolution than chronological time. We therefore used these values to also identify nonlinear dynamic eQTLs, whose effects may be present only at intermediate stages of the differentiation (**Fig. 3C**). We identified 74 nonlinear dynamic eQTL variants for the cardiomyocyte lineage (q<0.05), and 147 for the cardiac fibroblast lineage. Our time course study design is particularly useful for detecting transient nonlinear genetic effects which may not be found by studying only the initial or terminal cell types of a dynamic process such as differentiation.

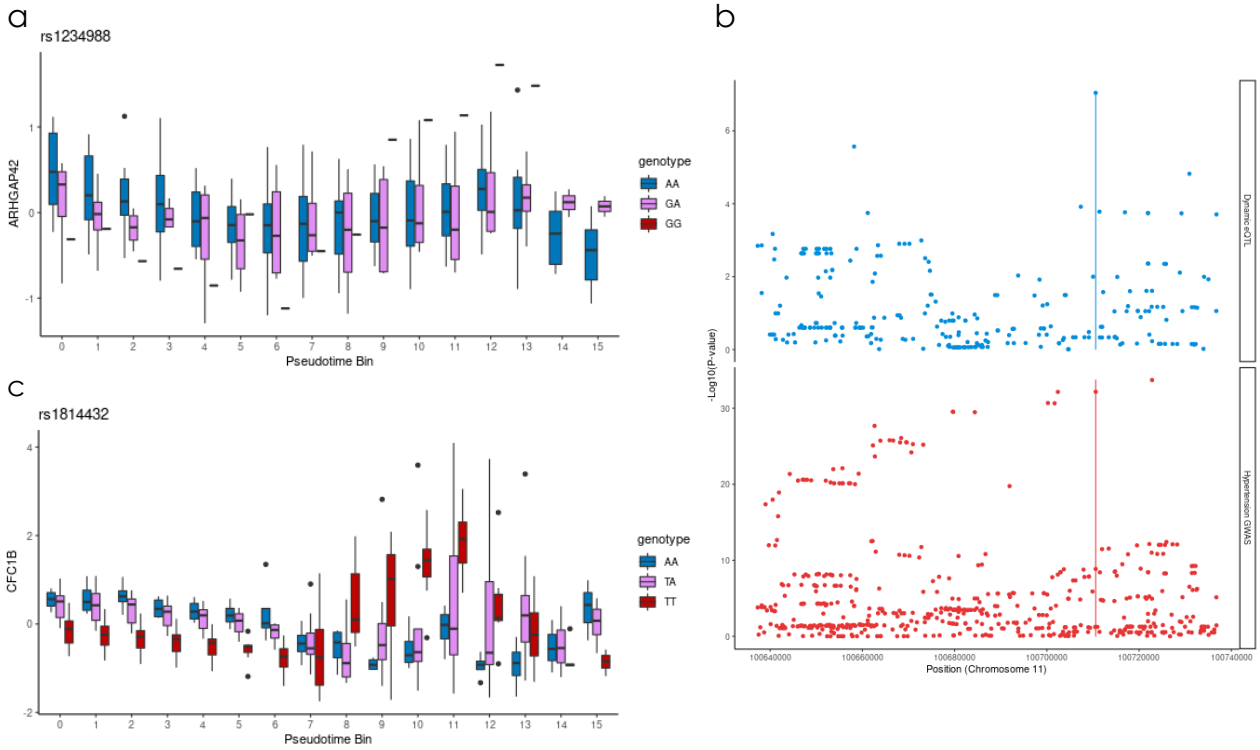


Fig. 3. Linear and nonlinear dynamic eQTLs. (A) rs1234988 is a linear dynamic eQTL for *ARHGAP42*; the effect of genotype (color) on *ARHGAP42* expression (y-axis) varies across pseudotime (x-axis). (B) A previously reported genome-wide association study (bottom) showed that hypertension is associated with genotype at the rs1234988 locus, where a dynamic eQTL for *ARHGAP42* was identified. (C) rs1814432 is a nonlinear dynamic eQTL for the gene *CFC1B*.

We examined the extent to which the dynamic eQTLs detected overlapped with eQTLs previously identified in GTEx (GTEx Consortium 2020). After subsetting to gene-variant pairs that were tested in both our data and GTEx, we found that the greatest replication of pseudotime-binned, cardiomyocyte lineage linear dynamic eQTLs occurred in atrial appendage tissue (π_1=0.50), while the greatest replication of pseudotime-binned, cardiac fibroblast linear dynamic eQTLs (as well as bulk) occurred in cultured fibroblasts (π_1=0.47, 0.56 respectively). However, by searching directly for dynamic effects across tissues rather than within a single tissue in isolation, we additionally identify eQTLs which were not found to be a significant eQTL in any tissue in GTEx. After subsetting to variant-gene pairs that were tested in both our data and GTEx, we found that 100 of the 359 (28%) linear dynamic eQTLs in the cardiomyocyte lineage were identified as eQTLs in GTEx. Similarly, only 22 of 75 (29.3%) nonlinear dynamic eQTLs on the cardiomyocyte lineage where previously identified as eQTLs in GTEx.

Deconvolution of bulk RNA sequencing data assigns lineage specificity to dynamic eQTLs

The information about the landscape of cardiomyocyte information obtained through single-cell RNA sequencing can also be applied retroactively to improve dynamic eQTL calling in bulk data. For each cell type that we identified in the single cell data, we computed a signature expression profile across the top 300 differentially expressed genes that were also measured in bulk (Methods). We then used CIBERSORTx to deconvolve our bulk data, assigning to each bulk sample a vector of cell type proportions (**Fig. 4A**, Newman et al. 2019). Deconvolution reveals that cell type heterogeneity is prominent between samples, particularly in days 7-15. This heterogeneity emphasizes the need to account for cell type proportion in measuring genetic regulatory dynamics, as these broad differences between cell lines can drive false positive associations between time and any genotype that is correlated with broad cell type proportion differences between cell lines.

We then used these cell type proportions to identify cell type specific effects, based on cell type interaction eQTLs (**ieQTLs**) for each known cell type that was observed in the single cell data (**Fig. 4B**). In this context, where cell types represent sequential steps along a developmental lineage, ieQTL calling is analogous to dynamic eQTL calling, using cell type proportion as a proxy for differentiation progress instead of time or pseudotime. Thus, ieQTLs for a cell type at an endpoint of the differentiation (iPSC, cardiomyocyte [CM], and cardiac fibroblast [CF]) are analogous to linear dynamic eQTLs, with additional information gained by assigning lineage specificity. CM and CF ieQTLs called with this approach were replicated in the previously used dynamic eQTL calling framework on the same bulk dataset (π_1=0.84 and 0.43, respectively). They additionally showed enrichment for genes related to myogenesis that had not been observed among bulk dynamic eQTLs (p=7e-4, both CM and CF ieQTL, compared to p=0.17, bulk dynamic eQTL). Notably, many of the CM- and CF-ieQTLs are lineage-specific, including some which are

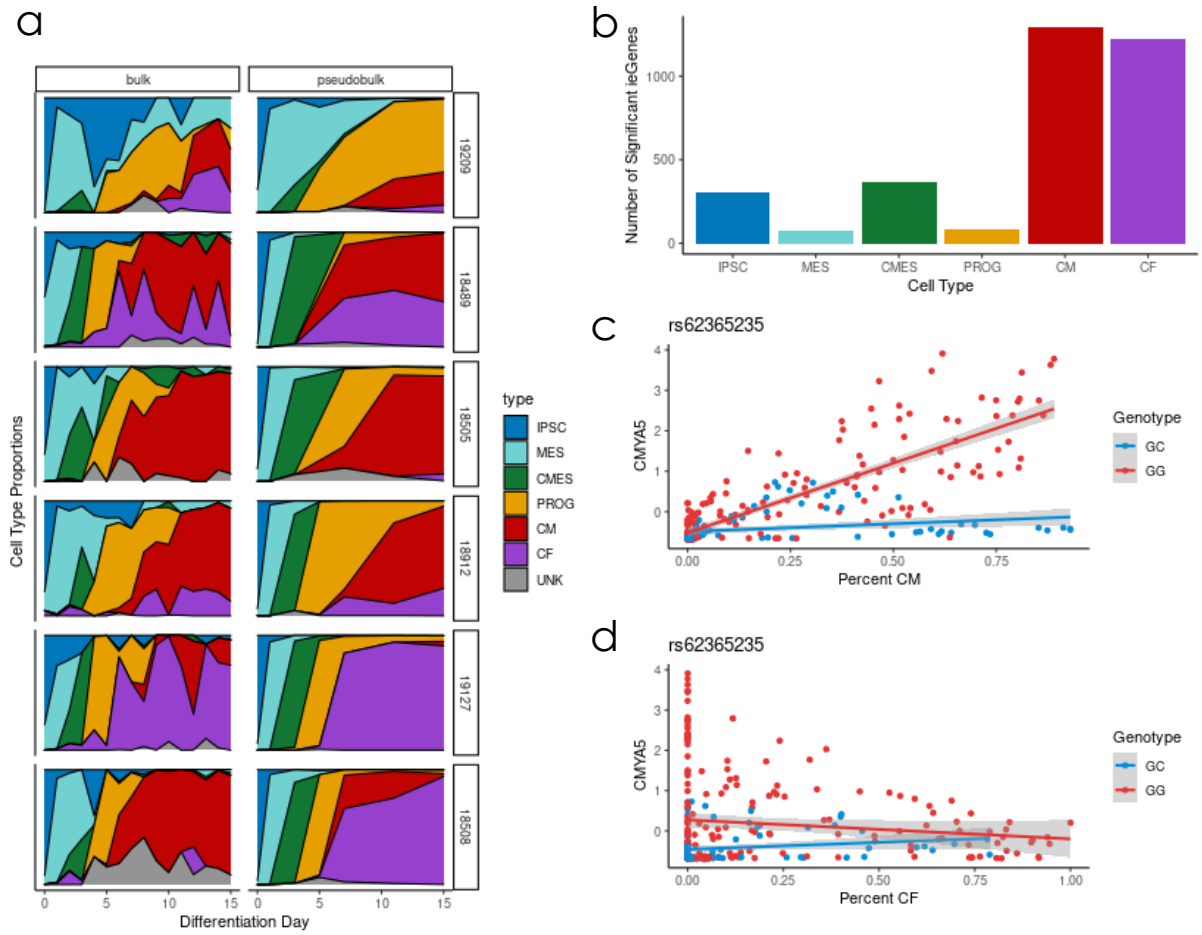


Fig. 4. Cell type deconvolution and interaction eQTL calling. (A) Cell type deconvolution was applied to decompose RNA expression of a mixed sample, aggregated over multiple cell types, into its constituent cell type proportions (Methods). Each row represents a cell line, collected in two separate experiments. In the left column, bulk RNA-sequencing data was collected for 15 timepoints (time on x-axis). In the right column, pseudobulk was aggregated across cells collected for 7 time points (time on x-axis). For pseudobulk data, deconvolution is not needed, as each cell is assigned to a cell type. Thus, "ground truth" cell type fractions are accessible as reflected here. (B) Number of genes with a cell type interaction eQTL in bulk for each of six cell types. (C-D) *CMYA5* has an interaction eQTL for the cardiomyocyte lineage (C) that is not identified in the cardiac fibroblast lineage (D).

potentially relevant to heart-related disease. **Fig. 4C-D** show an example of a cardiomyocyte interaction eQTL for cardiomyopathy-associated protein 5 (*CMYA5*), a gene which is highly expressed in heart and skeletal muscle and has previously been associated with cardiac hypertrophy (Nakagami et al. 2007). This variant was not previously identified by GTEx as an eQTL for *CMYA5*.

The discovery of additional dynamic eQTLs in bulk data with fewer differentiation time points sampled, as well as the ability to distinguish lineage-specific dynamic eQTLs from bulk data, demonstrate the utility of single-cell RNA-seq data matched to bulk samples to uncover dynamic genetic effects throughout a differentiation time course.

Discussion

Using iPSCs and their derived terminal cell types, we can identify genetic effects related to dynamic changes in gene expression over time. We used single-cell gene expression data to investigate the effects of gene regulatory and cell type composition changes throughout a cardiomyocyte differentiation time course. Single-cell data enables us to identify cells going down distinct differentiation trajectories, and to deconvolve heterogeneous cell types in matched bulk samples.

One question that arises from these single-cell data 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 through a semi-supervised pipeline which includes dimensionality reduction, unsupervised clustering, and visualization of expression patterns for known marker genes (**Fig. 1B**). Cells broadly express either of these gene sets in a mutually exclusive manner, suggesting that these gene sets represent two distinct cell types. The focus of this project was not to fully characterize these cell types, but instead to disentangle the broad effects of cell line differences in differentiation rate/ lineage preference from the dynamics of cis-regulation of gene expression. Still, the identity of these terminal cell types and the circumstances in which each trajectory might be favored is an interesting question.

These data suggest that there are differences in gene expression trajectory and ultimate cell fate that may arise in response to the same differentiation protocol. The identity of these terminal cell types, and the factors that might cause a cell line to favor one differentiation trajectory and ultimate cell type at the expense of another, are questions that have been explored in previous studies. In a study by D'Antonio-Chronowska et al. (2019), embryonic stem cell lines undergoing cardiac differentiation resulted in a heterogeneous cell type population. These cells were identified as either true cardiomyocytes --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 demonstrated that these two cell types were present in varying proportions in each individual cell line, 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).

The cardiomyocyte and epicardium framework explored by D'Antonio-Chronowska et al. may be useful in understanding the distinct differentiation trajectories present in our cardiac differentiations. The terminal non-cardiomyocyte cells expressing *COL3A1* in these samples may represent an endothelial or cardiac fibroblast cell type, which derive from the epicardium cell lineage. Cardiac fibroblasts express gene markers such as collagen and vimentin, which were found to be expressed in the terminal cells of this differentiation trajectory (Brade et al. 2013, Ieda et al. 2009, Zhang et al. 2019). The gene expression profile of *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 a reinforcement 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 with multiple replicates of each cell line, and compare the differentiation trajectories between these replicates. The relatively high correlation between these single-cell RNA-seq samples compared to matched bulk RNA-seq samples of the same cell line (Strober et. al 2019) suggests that there may be genetic factors involved in this trajectory decision -- although more rigorous testing should be performed to investigate this claim. We may also investigate whether subtle systematic differences exist between cell lines even in the iPSC stage (Day 0), 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 and Jerber 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 differentiation trajectories even before differentiation begins. This is an intriguing possibility, and more work should be performed to investigate whether the cell lines used here also exhibit distinct gene expression profiles early on that may correlate with the outcomes of any subsequent differentiation.

It is worth noting that the task of regression on an estimated latent variable (such as pseudotime, in dynamic eQTL calling, or cell type proportion, in cell type interaction eQTL calling), while biologically interesting, poses a challenge for statistical inference. Pseudotime and cell type proportions are estimated from expression data, rather than being experimentally measured. As a result, this represents an example of 'double dipping', where we determine which hypotheses to test downstream of exploratory data analysis. Such contexts have motivated interesting recent work in selective inference to address inflated type I error rates (Taylor et al. 2015, Gao et al. 2020). The jackstraw procedure (Chung and Storey 2015) accounts for selective inference in the context of regression on a continuous latent variable, but its application to pseudotime inference in this case is infeasible, as the procedure depends on latent variable inference for each of many resampling iterations. It is also worth noting that both dynamic and cell type interaction eQTLs assess effects of the interaction of genotype (a measured variable) with a latent variable, rather than the latent variable itself, which may mitigate the inflation effects of double dipping. We demonstrate in simulation that the fixed-effect linear model used in this study was conservative in

the presence of multiple measurements per individual, and did not lead to type I error inflation (Methods, Fig. S18). As unsupervised and semi-supervised machine learning methods provide increasingly reliable estimates of biologically important latent variables such as pseudotime, this will become an increasingly important area for further statistical methods development.

All together, the results from this study demonstrate the benefit of using single-cell RNA-sequencing with a balanced time course study design to investigate dynamic gene regulatory differences between individuals during cellular differentiation. Single-cell data offers a high-resolution view of the landscape of differentiation, which we leveraged to infer pseudotime along multiple differentiation trajectories. By isolating axes of variation of cis-regulatory dynamics (pseudotime within a particular lineage, rather than chronological differentiation day), we were able to identify a greater number of dynamic eQTLs with less than half as many collection time points as previous efforts in bulk RNA-seq data. The dynamic eQTLs detected included variants which overlapped known GWAS hits, demonstrating the utility of this approach in identifying causal loci that underlie risk for development of disease. We also used this data to lend new utility to bulk RNA-seq datasets, by assigning lineage specificity to dynamic eQTLs through the use of cell type interaction eQTL calling. While further follow-up studies should be performed to validate the function of these genomic loci and their potential relevance to downstream phenotypes, the dynamic genetic effects identified in this study and the methodology used to identify them provide a resource for investigating mechanisms underlying important biological processes such as cellular differentiation and perturbation response.

Acknowledgements

We thank Natalia Gonzales for providing feedback on the manuscript, and the lab of Anindita Basu for their support with Drop-seq. **Funding:** Y.G. and A.B. were supported by NIH/NIGMS R01GM120167. R.E. was supported by the NIH MSTP Training Grant T32GM007281. J.P. was supported by NIH/NIGMS T32GM119998. K.R. was supported by NIH/NHLBI 5F31HL146171. The computational resources were provided by the University of Chicago Research Computing Center. **Author contributions:** Y.G. and A.B. conceived the study. R.E. performed the experiments with assistance from K.R., and K.B. J.P. and R.E. analyzed the data, with assistance from B.J.S. G.Q. performed selective inference simulations. All authors wrote the paper. Y.G. and A.B. supervised this project. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The fastq files as well as processed and unprocessed expression matrices have been deposited in NCBI's Gene Expression Omnibus (Barrett et al., 2005) and are accessible through GEO Series accession number GSE175634 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE175634). The code for this analysis is available on Github at https://github.com/jmp448/sc-dynamic-eqtl/.

Materials and Methods

Samples We used induced pluripotent stem cell (iPSC) lines from 19 individuals from the Yoruba HapMap population. These iPSC lines were reprogrammed from lymphoblastoid cell lines and characterized previously (Banovich et al. 2018). All 19 individuals were 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, Waltham, MA) with Essential 8 Medium (A1517001, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning, Corning, NY). 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, Cambridge, UK).

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, Bristol, UK). "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. 2013). 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, San Jose, CA, #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, Cheonan, Korea, DHC-N01-2).

Using a 125 um Drop-seq microfluidic device, single cells were captured in droplets along with a DNA barcoded bead (ChemGenes, Wilmington, MA, 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, Santa Clara, CA, #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. S9**).

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. We then used the single-cell demultiplexing software 'demuxlet' to assign to each cell a probability that the cell is a doublet (Kang et al. 2018; **Fig. S3, S4, S14**).

The following filter was applied to remove 21,725 rare genes (out of 60,668) from downstream analysis:

- Gene must be detected in at least 10 cells

The following filters were then applied to remove 330,750 low-quality cells (out of 564,362) for downstream analysis:

- Maximum doublet probability of 0.3 from demuxlet
- Unambiguous assignment of the cell to an individual by demuxlet (maintain cells not assigned to 'doublet_ambiguous')
- Maximum of 25% mitochondrial reads
- Minimum of 300 unique genes detected (of the genes that passed the previous filtering step)

Following these filtering steps, an additional 2,826 cells were removed whose feature or read counts were more than 4 standard deviations away from the median. This left a total of 230,786 cells and 38,943 genes for downstream analysis.

***Cell cycle correction and normalization of single-cell expression data with* Seurat** We used the Seurat workflow for cell cycle regression in differentiating. Each cell was assigned a score for G2/M phase and S phase according to marker gene expression, and the difference between these scores was regressed out during normalization. The data was then normalized using the SCTransform function in (Stuart et al. 2019, Hafemeister and Satija 2019), producing corrected counts, log-normalized corrected counts, Pearson residuals, and a set of highly variable features. The Pearson residuals of 1,000 highly variable features were scaled so that each gene had unit variance across all cells for downstream analysis.

***Dimensionality reduction and clustering with* scanpy** Dimensionality reduction, clustering and pseudotime were performed using the *scanpy* package (Wolf et al. 2018), following Seurat object to h5ad conversion via the *sceasy* package (Cakir et al. 2020). The scaled Pearson residuals from 1000 highly variable features were used to compute 50 principal components (PCs), which were then embedded into a 2D UMAP plot (Fig. 1A,1C). These 50 PCs were also used to produce a neighborhood graph, and Leiden clustering was performed at resolution 0.35 to produce the clusters shown in Fig. 1C. (Several clusters are merged into the unknown cell type, as described below).

Lineage specification and pseudotime inference Based on marker gene expression patterns (Fig. 1B), 6 of the 10 Leiden clusters were annotated with known cell types. To facilitate trajectory reconstruction, 3 outlier clusters with less than 5,000 cells were removed. Cluster 7 contained a group of cells which did not express marker genes for cardiomyocytes or progenitor cell types, and instead expressed a group of genes that are specifically expressed in hepatocytes, a cell type

stemming from the endoderm layer rather than the mesoderm layer. This small population of cells drove a significant amount of variation in the data (**Fig. S5**), making it difficult to properly resolve the mesoderm-specific lineages that were the focus of this project. For this reason, the cells assigned to one of the mesoderm-specific lineages (clusters 1-6) were isolated, log-normalized gene expression was re-centered and re-scaled, and PCA was re-run on specifically these cells to properly focus on the variation among the lineages of interest. The top 3 re-computed PCs were used to calculate a new neighborhood graph, which was used to compute a new embedding to visualize specifically the two cardiac-related differentiating lineages (Fig. 2A). The bifurcation into separate cardiac fibroblast and cardiomyocyte lineages can clearly be observed in the PAGA plot (**Fig. S6)**, which was created with the previously described cell type annotations, the re-computed neighborhood graph, and an edge weight threshold of 0.15. This PAGA embedding was used to define the two lineages used for downstream lineage isolation tasks, where all iPSC, mesoderm, cardiac mesoderm, and cardiac progenitor cells are assigned jointly to both lineages, while cardiomyocyte and cardiac fibroblast (terminal cell types) are unique to their corresponding lineage. Finally, four diffusion components were computed from the new neighborhood graph, and diffusion pseudotime was used to assign pseudotime values to cells from both cardiac lineages.

Pseudobulk expression aggregation and normalization Although the noisiness of single cell expression profiles necessitates aggregation across cells before dynamic eQTL calling, an improved understanding of the differentiation landscape allows us to pursue an aggregation strategy that mitigates the confounding impact of cellular composition differences and offers greater power than dynamic eQTL calling on bulk samples. Three pseudobulk aggregation schemes were used in this study:

1. *Chronological differentiation day binning* - This strategy is most directly comparable to bulk RNA-sequencing. Aggregation is performed by taking the sum of SCTransform-corrected counts from all cells from the same differentiation day and individual.
2. *Lineage subsetting* - Differentiation day binning was performed within each lineage separately. As evidenced by the PAGA graph, all cells up to the progenitor cell type (PROG) are assigned to both lineages, only cells from the terminal cell type (cardiomyocyte or cardiac fibroblast) are unique to one lineage or another.
3. *Lineage subsetting & pseudotime binning* - After lineage subsetting, cells are partitioned into 16 quantile bins according to pseudotime. We chose 16 bins in order to directly compare to our previous 16 time-point bulk experiment (see **Fig. S7**). Aggregation then consists of the sum of SCTransform-corrected counts from cells within the same cell line and pseudotime bin.

After pseudobulk aggregation, low-depth samples with library size less than 10,000 were filtered out. Remaining samples underwent TMM normalization with singleton pairing through the *edgeR* package so that expression could be compared across samples for dynamic eQTL calling (Robinson et al. 2010, Robinson and Oshlack 2010). We then transform the TMM-normalized counts into compute counts per million (CPM) for each sample, and apply log normalization (with the *edgeR* package, which uses an approach to pseudocount addition that is adapted for library size). These logCPM expression values are used for QTL calling.

Bulk expression normalization In order to properly compare bulk RNA-seq data to our pseudobulk data, we reprocessed the bulk data from a previous experiment in a way that is intended to most closely match the logcpm pseudobulk expression. For this reason, we used transcripts per million (TPM) instead of previously used reads per kilobase of transcript, per million mapped reads (RPKM). For each sample, we first divided each gene's counts by the length in kilobase to compute reads per kilobase (RPK), and then fed these adjusted expression values into the same normalization pipeline as was used for pseudobulk counts (which are not biased by gene length) - TMM normalization with singleton pairing and logCPM adjustment, with the *edgeR* package. Since the input was reads per kilobase rather than counts, this gives logTPM expression values for use in QTL calling.

Sample PCA To identify primary sources of variation between samples, we ran principal component analysis (PCA) on the gene expression matrix for pseudobulk data. The first principal component is correlated with differentiation time (**Fig. S8**). For the top 10 PCs, we calculated the percent variance explained of each principal component by each technical factor recorded during sample collection (**Fig. S9**).

Cell line collapsed PCA To perform dynamic eQTL calling, we search for changes in gene expression over time that are correlated with a specific genotype. This can be confounded by broad differences between cell lines across the differentiation time course, such as differences in differentiation speed, lineage preference, or technical factors. For example, assume cell lines with genotype G at locus *i* generally have increasing proportions of cardiomyocytes over time, while cell lines with genotype C at locus *i* have increased proportions of cardiac fibroblasts over time. In this case, any gene whose expression is upregulated in cardiomyocytes will appear to have a dynamic eQTL at locus *i*, regardless of any cis-regulatory dynamics related to that gene, which constitute the intended focus of this study.

With single-cell data, we are able to more directly account for some of these factors, namely differentiation speed (with pseudotime binning) and lineage preference (with lineage subsetting). However, it remains useful to control for any broad cell line differences in this more unsupervised fashion, as any broad effects could drive false positive QTL detection.

We used a “cell line collapsed PCA” approach to identify such patterns across the entire time course (Strober et al. 2019). To identify cell line collapsed PCs, we rearranged the gene expression matrix from the standard pseudobulk expression 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 to have zero mean and unit variance, we applied PCA to this matrix to learn a low dimensional representation. 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 both linear and nonlinear dynamic eQTLs, in both bulk and pseudobulk.

To detect cell line specific patterns that may potentially be confounding variables in our dynamic eQTLs, we calculated the frequency at which each pair of cell lines share the same genotype across all significant dynamic eQTLs, compared to what is expected by chance. After controlling for five cell line collapsed PCs, cell lines do not share the same genotype at more significant

eQTLs than expected by chance, confirming that cell line PCs adequately address these potential confounding effects (Fig. S15).

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). For eQTL analyses, we filtered to variants with no missingness and a minor allele frequency of at least 0.1 across the 19 individuals present.

Dynamic cis-eQTL test selection We selected which genes to check for dynamic eQTLs based on the following filters:

- Gene must have at least 0.1 CPM in at least 10 bulk/ pseudobulk samples
- Gene must have at least 6 counts (reads) in at least 10 samples

Both of these filters were applied separately for each aggregation scheme. We tested all variants within 50kb of the transcription start site of each gene. Transcription start sites were obtained from Gencode's release 37 (GRCh38.p13, Frankish et al. 2019) basic gene annotation, and matched to mapped genes by Ensembl gene ID. The total number of tests is presented alongside the number of dynamic eQTLs detected in tables 1 and 2.

Linear dynamic eQTLs using single-cell pseudobulk data 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).

Following the method used in Strober et al 2019, we built a separate linear model for each tested variant-gene pair. Specifically, let t denote the time point (or, for pseudotime binning, the median pseudotime value across cells constituting the pseudobulk sample) 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:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 PC_c^1 + \cdots + \beta_7 PC_c^5 + \beta_8 PC_c^1 t + \cdots + \beta_{12} PC_c^5 t + \beta_{13} G_c t, \sigma)$$

We used lmFit from the limma package to fit this model, and used a t-test to measure the significance of the genotype and time coefficient (β_{13}).

Bonferroni correction was applied to account for multiple SNPs being tested per gene, and Storey's q-value was used to control false discovery rates at the gene level, after selecting the most significant dynamic eQTL per gene. Genetic correlation among significant dynamic eQTLs, which could be indicative of broad effects driving inflated type I error rates, did not

appear to significantly differ from background variants within 50kb of a TSS matched for minor allele frequency (**Fig. S15**).

Nonlinear dynamic eQTLs using single-cell pseudobulk data 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:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 t + \beta_3 t^2 + \beta_4 PC_c^1 + \cdots + \beta_8 PC_c^5 + \beta_9 PC_c^1 t + \beta_{10} PC_c^1 t^2 \ldots + \beta_{17} PC_c^5 t + \beta_{18} PC_c^5 t^2 + \beta_{19} G_c t + \beta_{20} G_c t^2, \sigma)$$

Once again, time is either time of collection, or median pseudotime of the sample. As before, we used lmFit from the limma package to fit this model, and this time used a similar t-test to measure the significance of the genotype and quadratic time coefficient (β_{20}). Multiple testing correction was applied as with linear dynamic eQTL calling.

Permutation analysis We assessed calibration of our dynamic eQTL calling methods with permutations. If we permute the time variable in the interaction term, we do not expect this term to properly capture interactions between genotype and time. For each variant-gene pair, we performed an independent permutation of the time variable in the interaction term, across all (cell line, day) samples. The results of this analysis are shown in **Fig. S10**. As another check for confounding factors, we checked whether dynamic eQTLs were enriched for genotypes shared between any particular pair of individuals (suggesting broad individual differences could be driving the dynamic eQTLs, **Fig. S10**).

Simulations to examine type I errors due to 'double dipping' We conducted simulations to evaluate potential type I error inflation caused by selective inference. We simulated gene expression data from the following linear mixed model:

$$Y_{ijk} = \beta_k G_{ik} + \alpha_k M_{ij} + a_{ik} + \epsilon_{ijk},$$

Here Y_{ijk} is the expression of gene k in cell j of individual i, where $k = 1, \ldots, 1000$, $j = 1, \ldots, 100$ and $i = 1, \ldots, n$. The sample size n is 10 or 20. We assumed one cis-eQTL per gene. To simulate the genotype G_{ik}, we first generated the minor allele frequency (MAF_k) from $Uniform(0.1, 0.5)$ and then generated $G_{ik} \sim binomial(2, MAF_k)$.
The other variables included genetic effect size β_k, cell maturity M_{ij} and its effect size α_k, individual-specific random effect a_{ik} and error term ϵ_{ijk}. They were generated from the following distributions:

$$\beta_k \sim N(0, \sigma_\beta^2), \qquad M_{ij} \sim N(0,1), \qquad \alpha_k \sim N(0, \sigma_\alpha^2)$$

$$(a_{i1}, \ldots, a_{i,1000}) \sim N(0, \sigma^2 \Sigma), \qquad (\epsilon_{ij1}, \ldots, \epsilon_{ij,1000}) \sim N(0, \sigma_e^2 \Sigma),$$

Note that $(a_{i1}, \ldots, a_{i,1000})$ are i.i.d. across individuals and $(\epsilon_{ij1}, \ldots, \epsilon_{ij,1000})$ are i.i.d. across individuals and cells, but they are both correlated across genes. To construct a realistic correlation structure, we chose Σ to be the correlation matrix of the expression of 1000 randomly

selected genes from our pseudo bulk data. We fixed $\sigma_\alpha^2 + \sigma^2 = 0.3$ so that cell maturity and individual specific random effect explained 30% variance of expression and varied $\frac{\sigma_\alpha^2}{\sigma^2} = 0, 0.1, 0.5, 1, 2, 10$. We then generated the genetic effect size $\beta_k \sim N(0, 0.1^2)$ or $N(0, 0.4^2)$, corresponding to on average 0.4% or 6.3% variance of gene expression explained by genetic effects. The variance of the error term σ_e^2 was chosen so that the expression of each gene has unit variance.

We defined the pseudo time in this simulation study to be the first gene expression principal component (PC). We divided the cells into three equal pseudo time bins and averaged expression of the cells for each individual in each pseudo time bin into pseudo bulk expression ($\tilde{Y}_{ilk}$). We also calculated the average pseudo time for cells within each pseudo bulk sample, denoted by t_{il}. We tested two models for dynamic eQTL calling (fitted for each gene k separately): 1) linear mixed model with individual-specific random effects $\tilde{Y}_{ilk} \sim G_{ik} + t_{il} + G_{ik}t_{il} + (1|\text{individual})$; 2) linear model $\tilde{Y}_{ilk} \sim G_{ik} + t_{il} + G_{ik}t_{il}$ without random effects. Type I error was calculated across 1000 genes (**Fig. S18**). The simulation suggests that a fixed-effect linear model for dynamic eQTL calling, as used in this study, was conservative in the presence of multiple measurements per individual and did not lead to type I error inflation. The more powerful linear mixed model did lead to moderate inflation.

Correlation between bulk and pseudobulk data We calculated the Pearson correlation of the normalized gene expression matrix from matched bulk RNA-seq data (Strober et al 2019) with the normalized gene expression matrix from pseudobulk RNA-seq data. We observed a high correlation of gene expression values between bulk and pseudobulk samples of any given differentiation day (**Fig. S11**), and a consistent pattern of correlation for all cell lines (**Fig. S12**).

Bulk dataset deconvolution using single cell data Cell type deconvolution was performed using CIBERSORTx (Rusk 2019). The method was first assessed for accuracy using pseudobulk data, where a ground truth is available. Cells from each annotated cell type were split into training (60% of cells) and testing (40%) groups. The annotated Seurat object was subset to training data, and the *FindAllMarkers* command was used to identify a subset of 404 genes for use in deconvolution. We removed genes that were not measured in bulk, leaving 317 genes for use in deconvolution. A gene expression signature matrix was created from exclusively the training data by taking the sum of SCTransform-corrected counts within each cell type. Normalization of the signature matrix was performed using edgeR: normalization factors were first computed with 'TMMwsp' method, then TMM-normalized counts were converted to counts per million. To assess the accuracy of this approach, we then used the same normalization pipeline to aggregate pseudobulk by sample for the testing data, where samples corresponded to a (cell line, differentiation day) combination (**Fig. S13**). To perform deconvolution of the bulk RNA sequencing data, we used the signature matrix described above and subset the bulk data to the 317 genes contained in the signature matrix.

Cell type interaction eQTLs To account for variable cell type composition in bulk RNA-seq data, rather than looking for cis-eQTLs whose effects are modulated by time (linear dynamic eQTLs), we looked for those whose effects are modulated by cell type proportion (Kim-Hellmuth et al. 2020). This mitigates the confounding impact of lineage preference on dynamic eQTL calling, as well as differences in differentiation speed (to the extent that this is captured by

cell type proportion). To do so, we replaced the time variable in the dynamic eQTL model with cell type proportion as follows:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 K_{ct} + \beta_3 PC_c^1 + \cdots + \beta_7 PC_c^5 + \beta_8 PC_c^1 K_{ct} + \cdots + \beta_{12} PC_c^5 K_{ct} + \beta_{13} G_c K_{ct}, \sigma)$$

Where K_{ct} is the CIBERSORTx inferred cell type proportions in the sample. Separate models were built for each variant-gene pair, in each cell type except the 'unknown' cell type. We additionally explored a model in which we regressed out all cell type proportions (except the unknown cell type, as cell type proportions are constrained to sum to 1).

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 K_{IPSC} + \ldots + \beta_7 K_{CM} + \beta_8 PC_c^1 + \cdots + \beta_{12} PC_c^5 + \beta_{13} PC_c^1 K_{ct} + \cdots + \beta_{17} PC_c^5 K_{ct} + \beta_{18} G_c K_{ct}, \sigma)$$

Note that while all fixed cell type proportion terms are included as covariates, there is only one interaction term for a single cell type proportion. Therefore, once again, separate models were fit for each variant-gene pair, in each cell type except 'unknown'. We found that regressing out additional cell types, not just the one included in the interaction term, led to detection of a greater number of genes with a cell type interaction eQTL (**Fig. S16**). To check whether these additional covariates were in fact introducing false positive associations between individuals, we measured the pairwise genetic correlation between cell lines among the top hits detected after regressing out additional cell type proportions. We then compared this to the genetic correlation among a set of hits detected before regressing out additional cell type proportions, matched for minor allele frequency. We did not see an increase in genetic correlation among significant tests introduced by incorporation of additional covariates (**Fig. S17**). However, we did observe a lower replication rate of this expanded set of interaction eQTLs among linear dynamic eQTLs (π_1=0.69 and 0.32, respectively, compared to 0.84 and 0.43 under the first model).

We also explored including sample-level principal components as covariates in the linear model:

$$E_{ct} \sim N(\mu + \beta_1 G_c + \beta_2 U^1 + \ldots + \beta_7 U^5 + \beta_8 PC_c^1 + \cdots + \beta_{12} PC_c^5 + \beta_{13} PC_c^1 K_{ct} + \cdots + \beta_{17} PC_c^5 K_{ct} + \beta_{18} G_c K_{ct}, \sigma)$$

Where U^1 represents the first sample principal component, as opposed to PC_c^1, the first cell line principal component. Here, we again found that additional covariates led to an increased number of cell type interaction eQTLs detected (**Fig. S16**): for several cell types (pluripotent cells, mesoderm and progenitor) this figure continued to increase with up to 30 principal components regressed out. With the terminal cell types where more interaction eGenes were detected, the maximum number of hits detected occurred after regressing out 10 principal components. The replication rate among dynamic eQTLs decreased as the number of hits detected increased (π_1=0.63 and 0.30 for cardiomyocyte and cardiac fibroblast, respectively, after 5 PCs were regressed out; 0.59 and 0.38 after 10; 0.64 and 0.42 after 20; 0.68 and 0.44 after 30). The results from fitting the first model are reflected in the main text.

Overlap with published GTEx eQTLs We used the GTEx v8 release to evaluate replication and overlap of our dynamic eQTLs with variants previously detected in adult tissues. To assess

861 replication in each tissue, we used the qvalue package in R (Storey 2003) to compute
862 π_1 replication rates among all variant-gene pairs that were declared dynamic eQTLs that were
863 also tested in GTEx. To determine the percentage of variant-gene pairs that were declared both
864 dynamic eQTLs and significant *cis* eQTLs in GTEx, we incorporated *cis* eQTLs from all tissues.
865

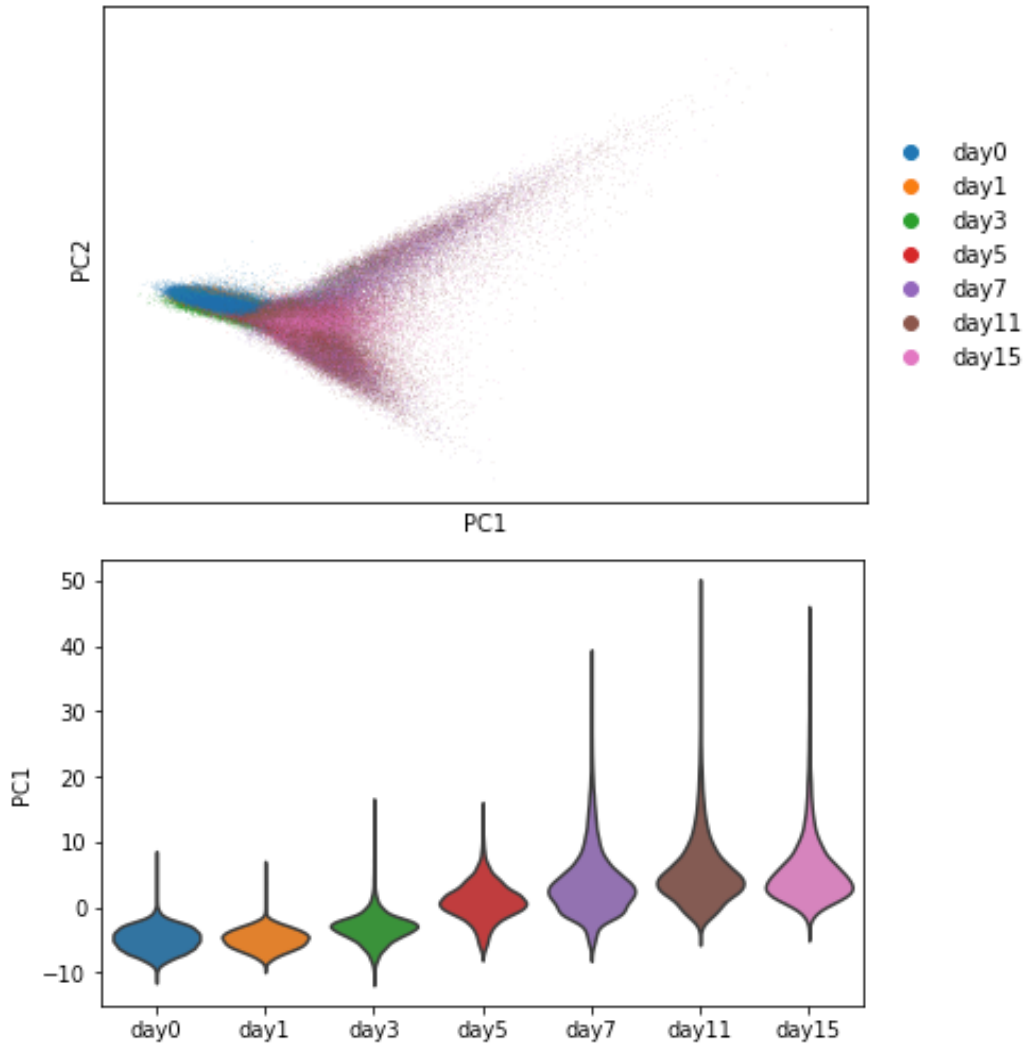


Fig. S1: Principal component analysis of single cell data. (Top) Principal components biplot for single cell data, colored by differentiation day. The first principal component is correlated with differentiation progress (Bottom), while the second principal component differentiates between the two terminal cell types.

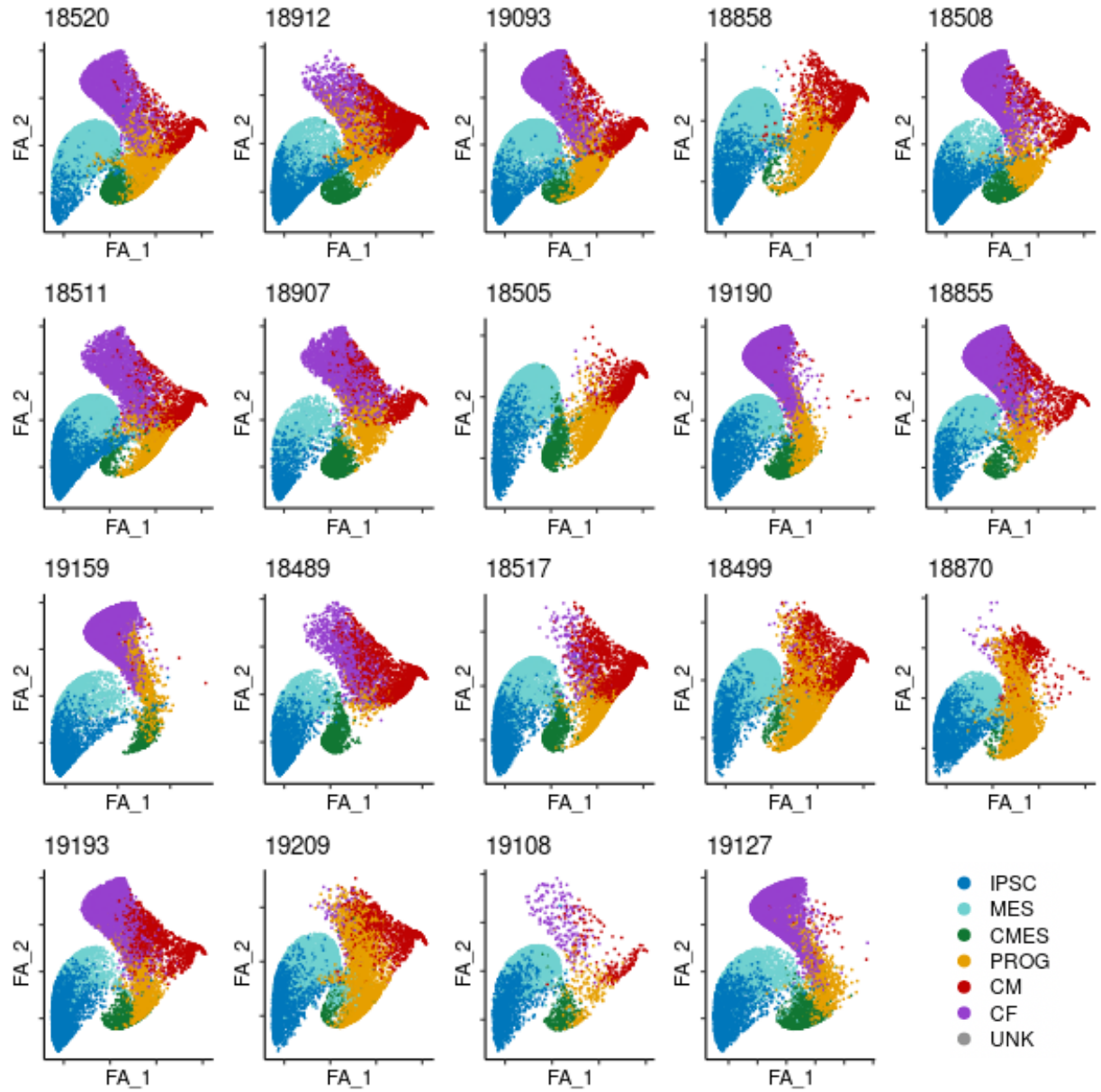


Fig. S2: Cell lines display differences in trajectory preference. The force atlas embedding which was learned from all cells jointly is shown for each individual cell line, colored by cell type.

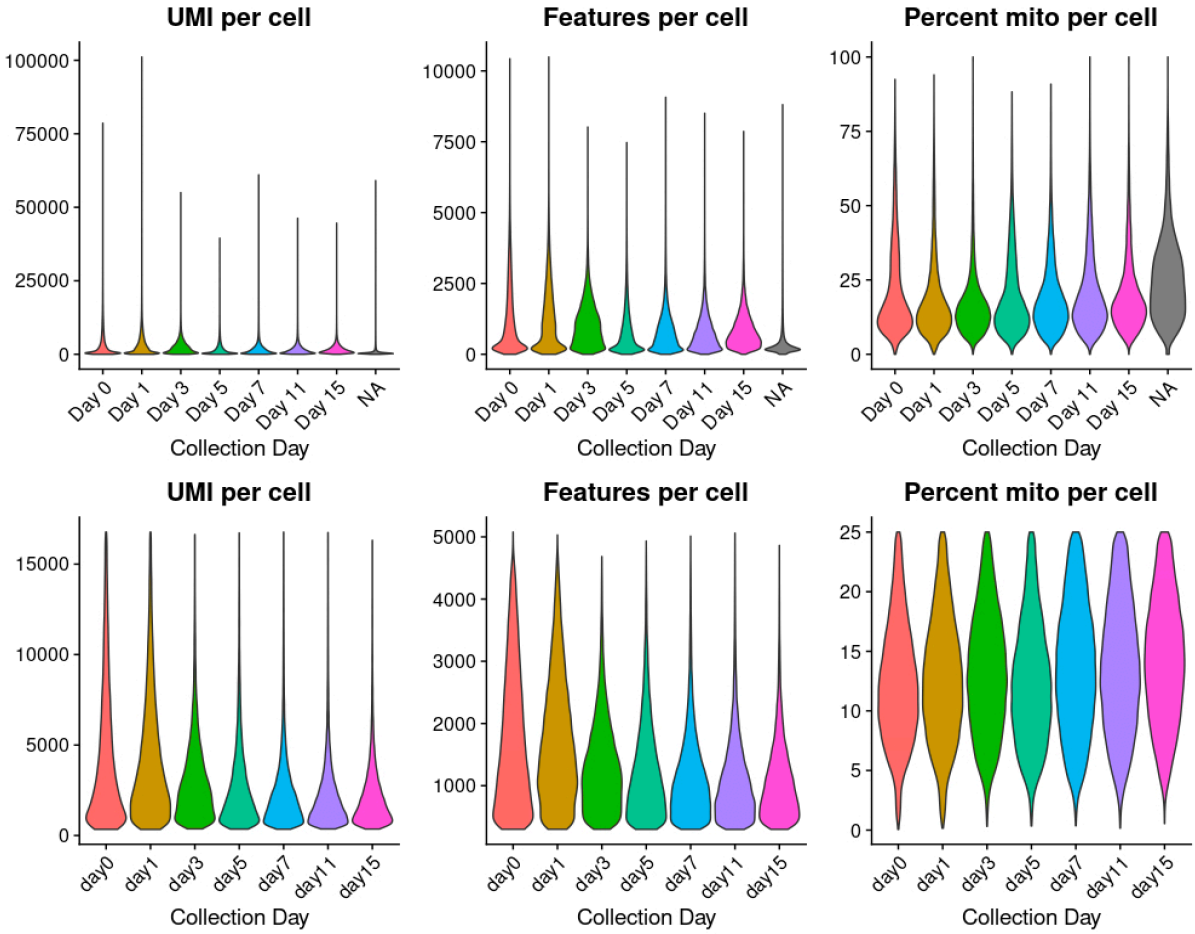


Fig. S3. Number of UMIs, genes, and percent mitochondrial reads per cell in single cell data, by day. Distribution of the number of Unique Molecular Identifiers (UMIs) per cell, number of genes per cell, and the percent mitochondrial reads per cell in full single cell dataset, prior to (top row) and after (bottom row) filtering as described in Methods (*RNA-seq quantification*). X-axis separated by differentiation day.

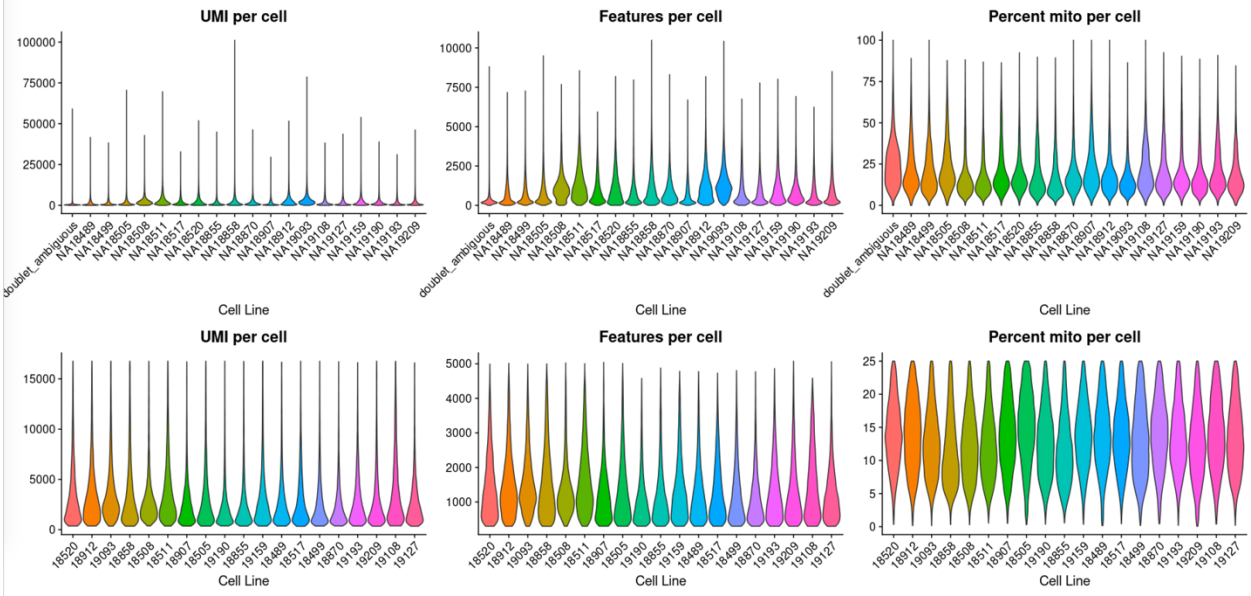


Fig. S4. Number of UMIs, genes, and percent mitochondrial reads per cell in single cell data, by individual. Distribution of the number of Unique Molecular Identifiers (UMIs) per cell, number of genes per cell, and the percent mitochondrial reads per cell in full single cell dataset, prior to (top row) and after (bottom row) filtering as described in Methods (*RNA-seq quantification*). X-axis separated by cell line.

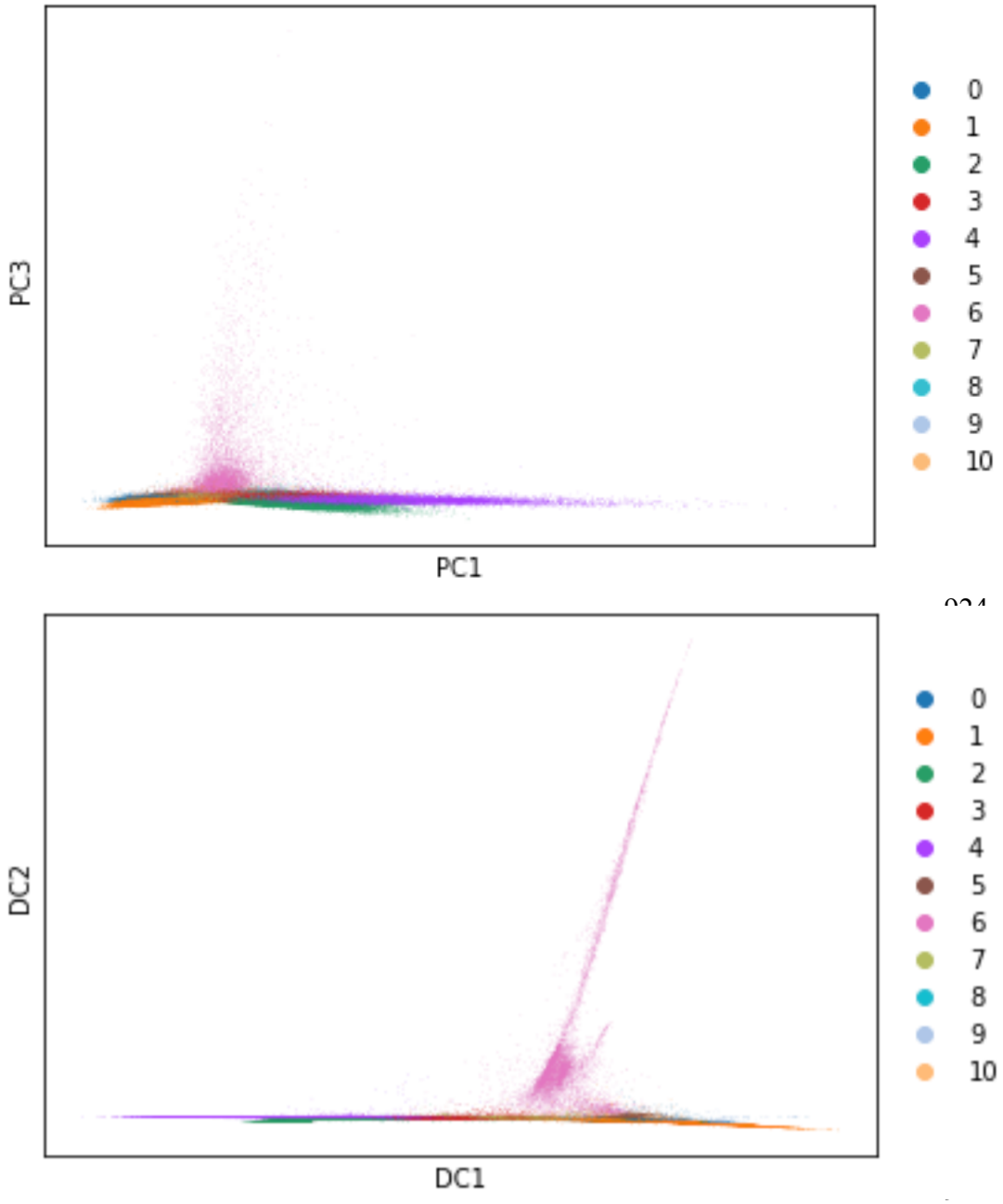


Fig. S5: Cell cluster 6 appears to be an outlier cluster. This group of cells which underexpresses cardiac markers from all stages of differentiation and overexpresses endoderm markers such as *APOA1* and *AFP* is picked up by the third principal component (top), and largely drives the variation behind the second diffusion component (bottom). The variation driven by relatively small population of cells interferes with reconstruction of biologically feasible trajectories, and was removed from downstream analysis.

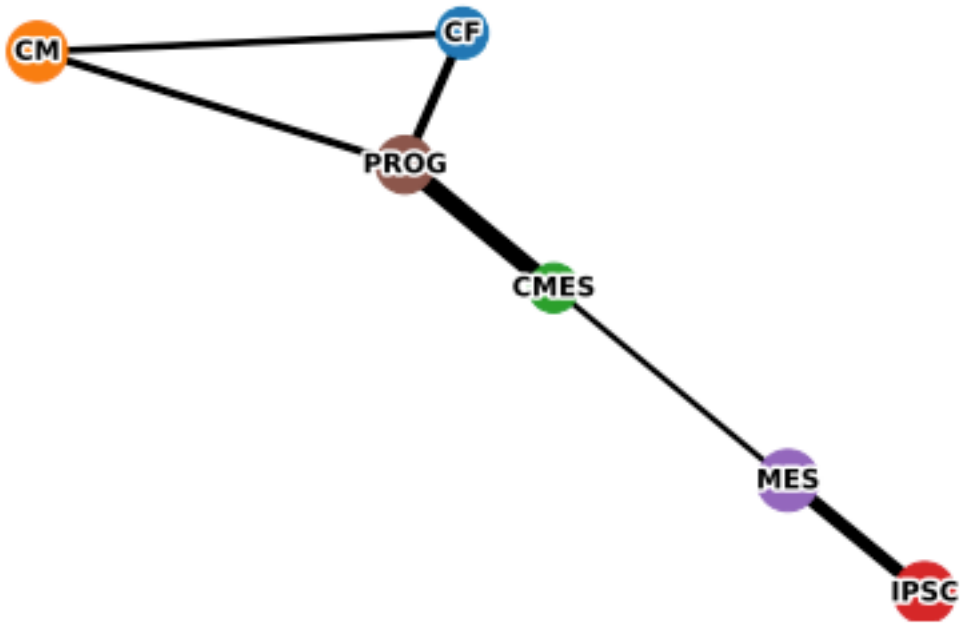


Fig. S6: PAGA identifies a bifurcation in cellular differentiation. PAGA identifies a bifurcation into cardiomyocyte and cardiac fibroblast cell types after the cardiac progenitor stage.

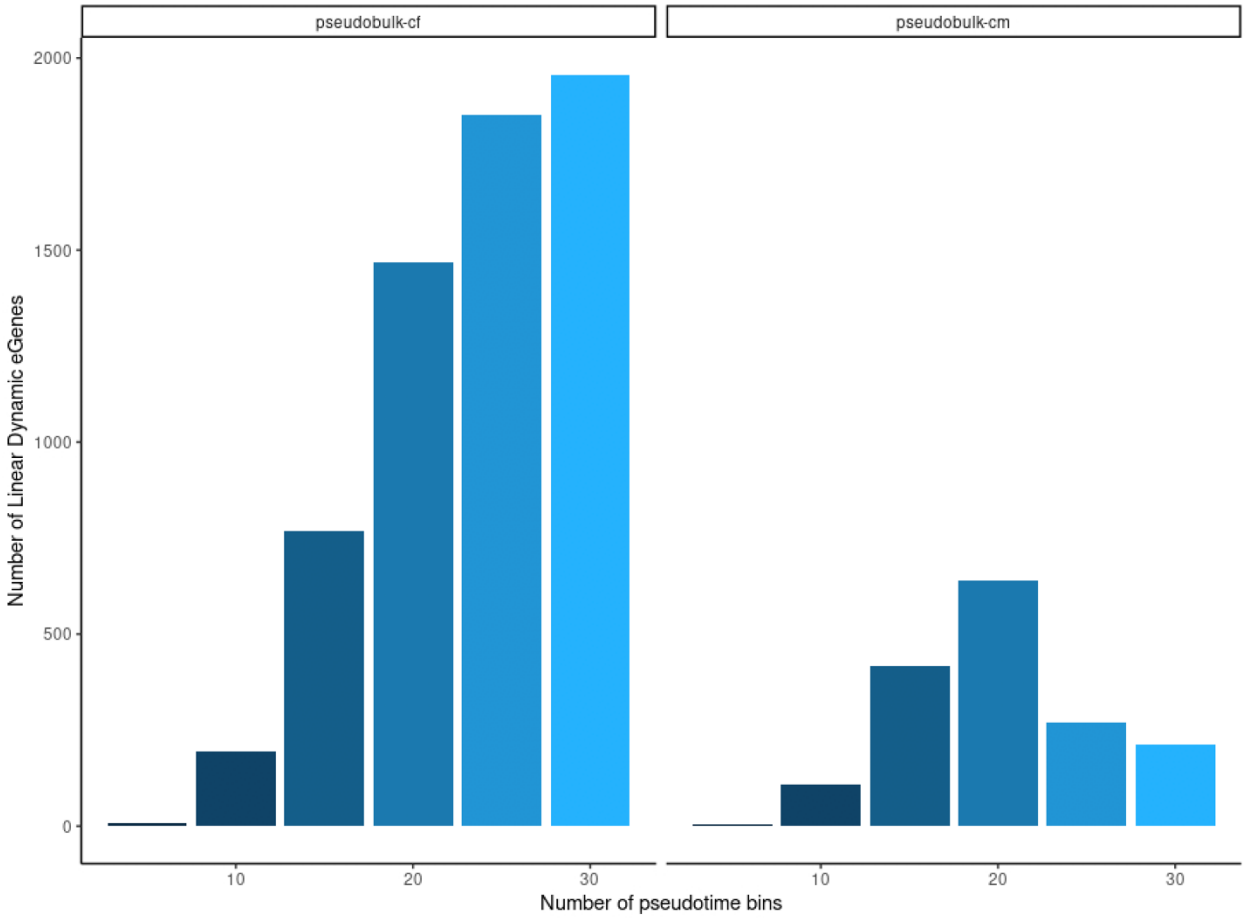


Fig. S7: Dynamic eQTL detection rates across multiple bin sizes. Y-axis shows the number of significant linear dynamic eGenes (genes with a dynamic eQTL, q<0.05) for a variety of numbers of pseudotime quantile bins (x-axis) for both the cardiac fibroblast (pseudobulk-cf, left) and cardiomyocyte (pseudobulk-cm, right) lineages.

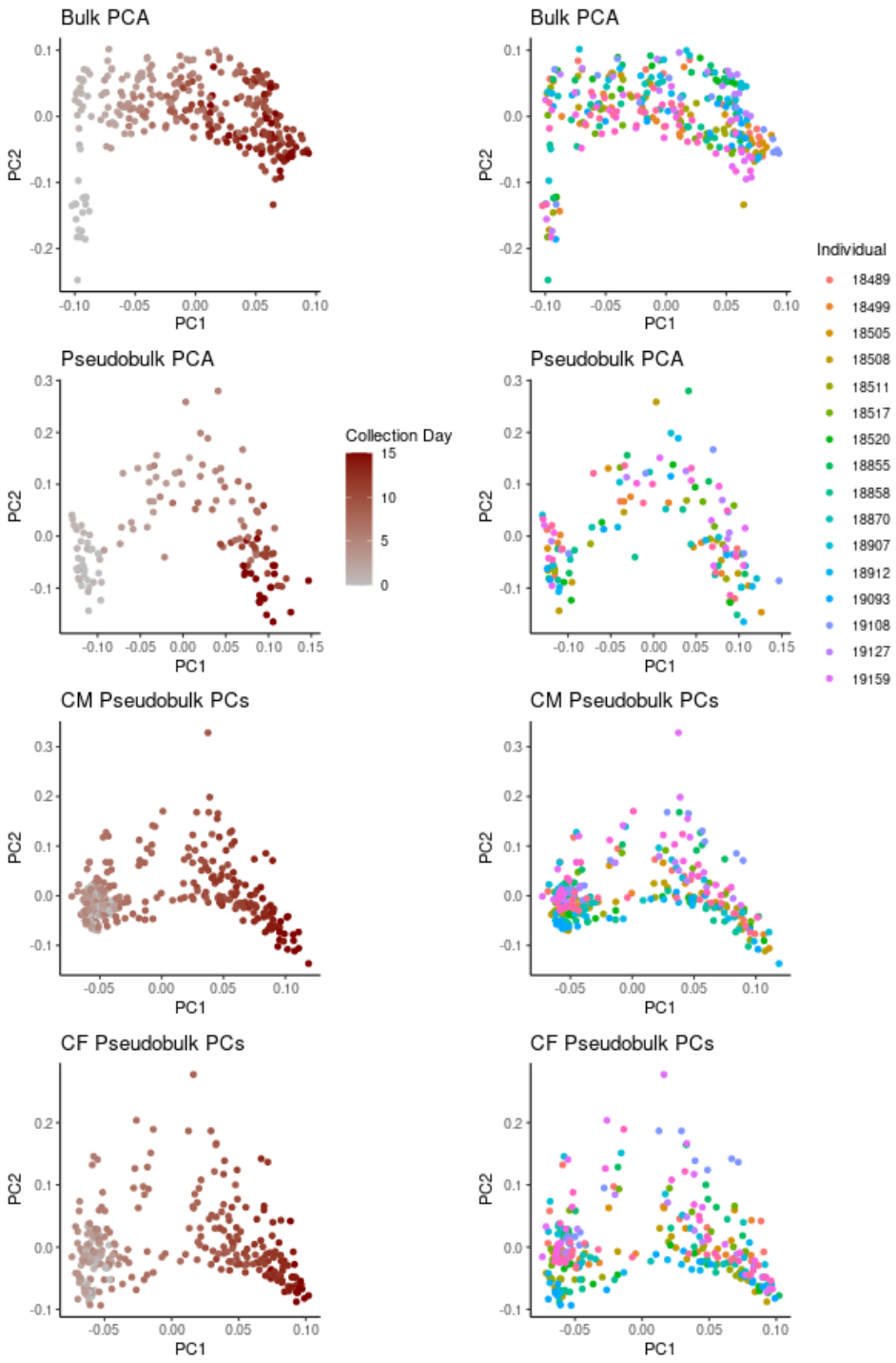
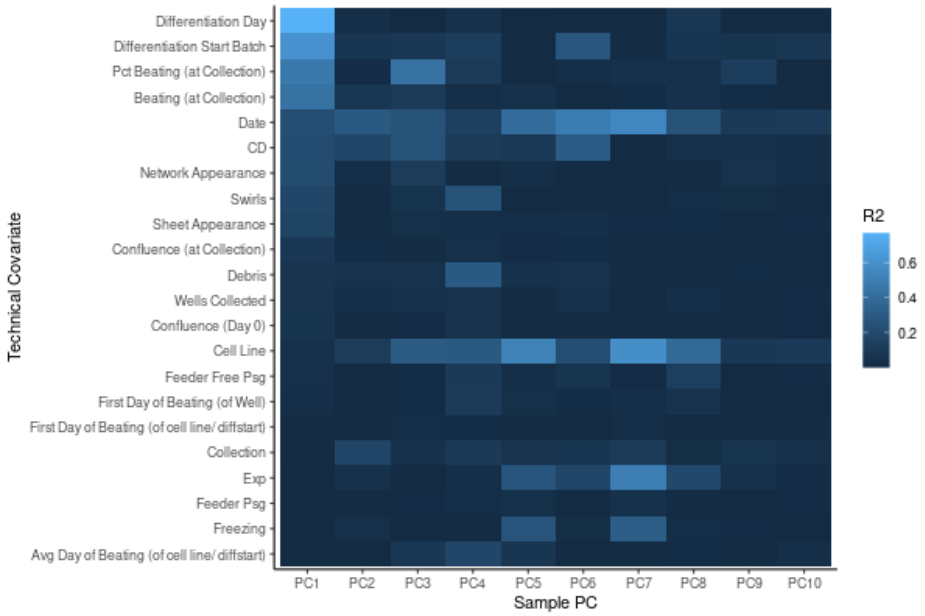


Fig. S8: PCA on pseudobulk and bulk samples identifies differentiation progress as primary source of variation. PCA on bulk (row 1), single cell data aggregated into pseudobulk by differentiation day / individual (row 2), cardiomyocyte lineage-specific single cell data aggregated into pseudobulk by pseudotime / individual (row 3), and cardiac fibroblast lineage-specific single cell data aggregated into pseudobulk by pseudotime / individual (row 4). Samples colored on a gradient by (left column) differentiation day or pseudotime bin, or (right column) cell line.



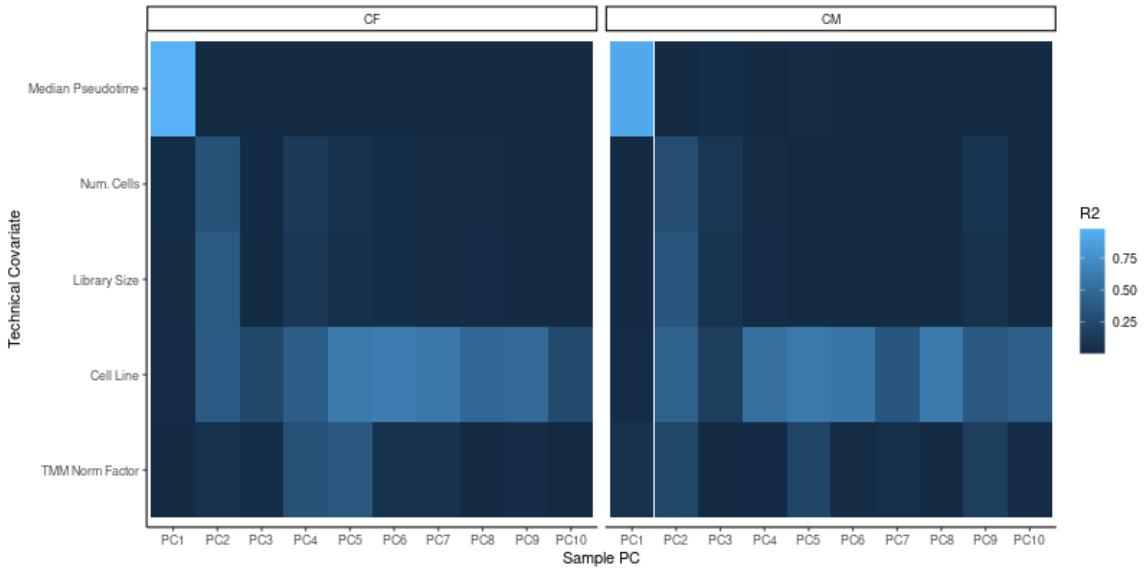


Fig. S9: PCs percent variance explained by technical factors in single cell data. (a) Variance explained of each gene expression principal component (1-10) for pseudobulk samples aggregated by cell line and differentiation day using recorded covariates, including: percent cells beating (visually assessed), differentiation day, collection day, culture confluence, cell morphology (visually assessed), and cellular debris. (b) Variance explained of principal components for pseudobulk samples aggregated by cell line pseudotime bin for cardiac fibroblast (CF, left) and cardiomyocyte (CM, right) lineages. Technical covariates shown are cell line, library size, median pseudotime, number of cells, and the normalization factor used for TMM normalization, from the edgeR package (see *Methods*).

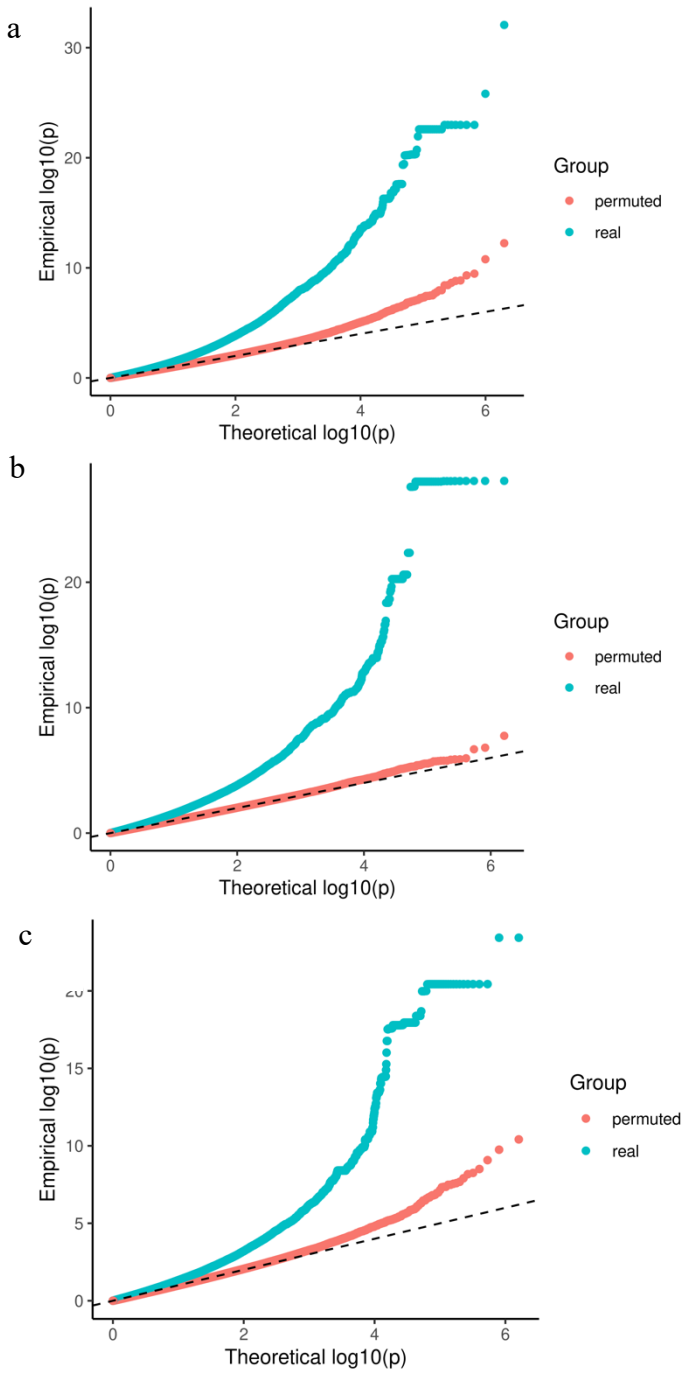


Fig. S10: Permutation analyses. Permutation analyses (see *Permutation analysis* in Methods) do not suggest inflation in bulk (a), pseudotime-binned cardiomyocyte-subset pseudobulk (b), or pseudotime-binned cardiac fibroblast-subset pseudobulk (c). The p-values from this study are shown in blue, while those obtained from a permutation test are shown in red.

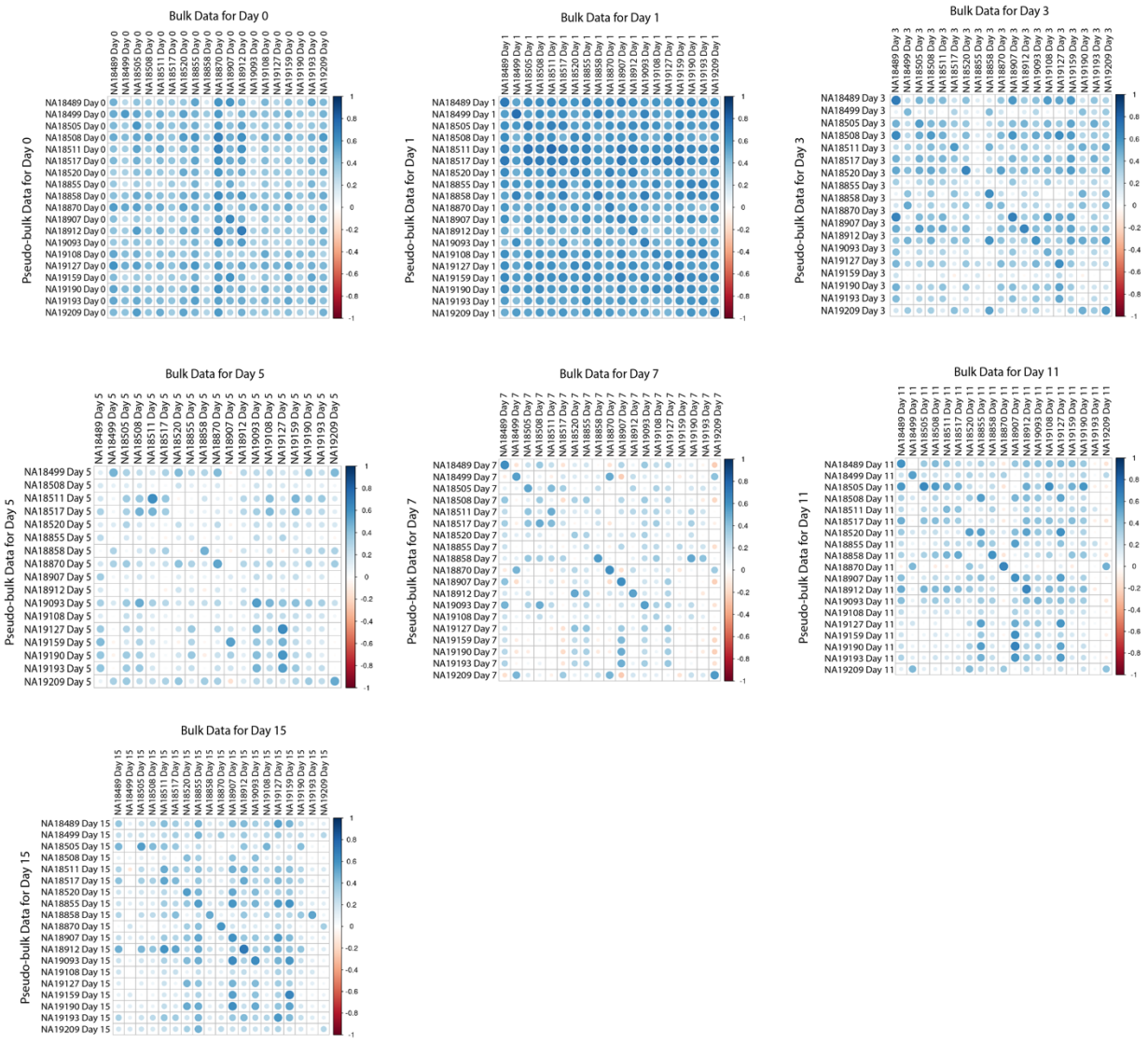


Fig. S11: Correlation of bulk and pseudobulk data by day. Pearson correlation between single-cell pseudobulk data and bulk RNA-seq data (Strober et al 2019) for each individual; panels separated by differentiation day.

Fig. S12: Correlation of bulk and pseudobulk data by individual. Pearson correlation between single-cell pseudobulk data and bulk RNA-seq data (Strober et al 2019) for each differentiation day; panels separated by individual.

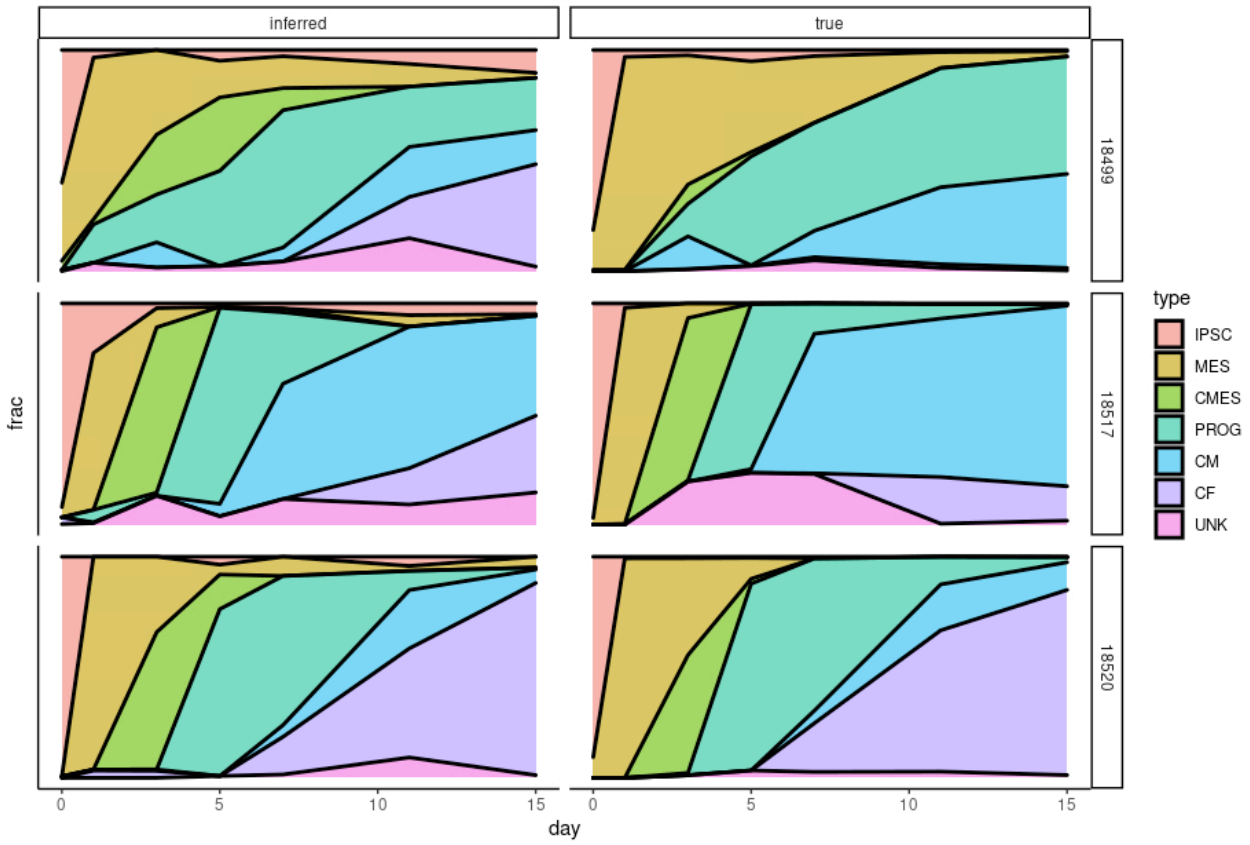


Fig. S13: CIBERSORTx assessment in pseudobulk. Assessment of CIBERSORTx performance in pseudobulk, where 'ground truth' is available. CIBERSORTx-estimated cell type proportions from differentiation day-binned pseudobulk data for three cell lines is shown at left ('inferred'), compared to true cell type proportions ('true', right), as determined by the cell type annotation approach described in the supplement.

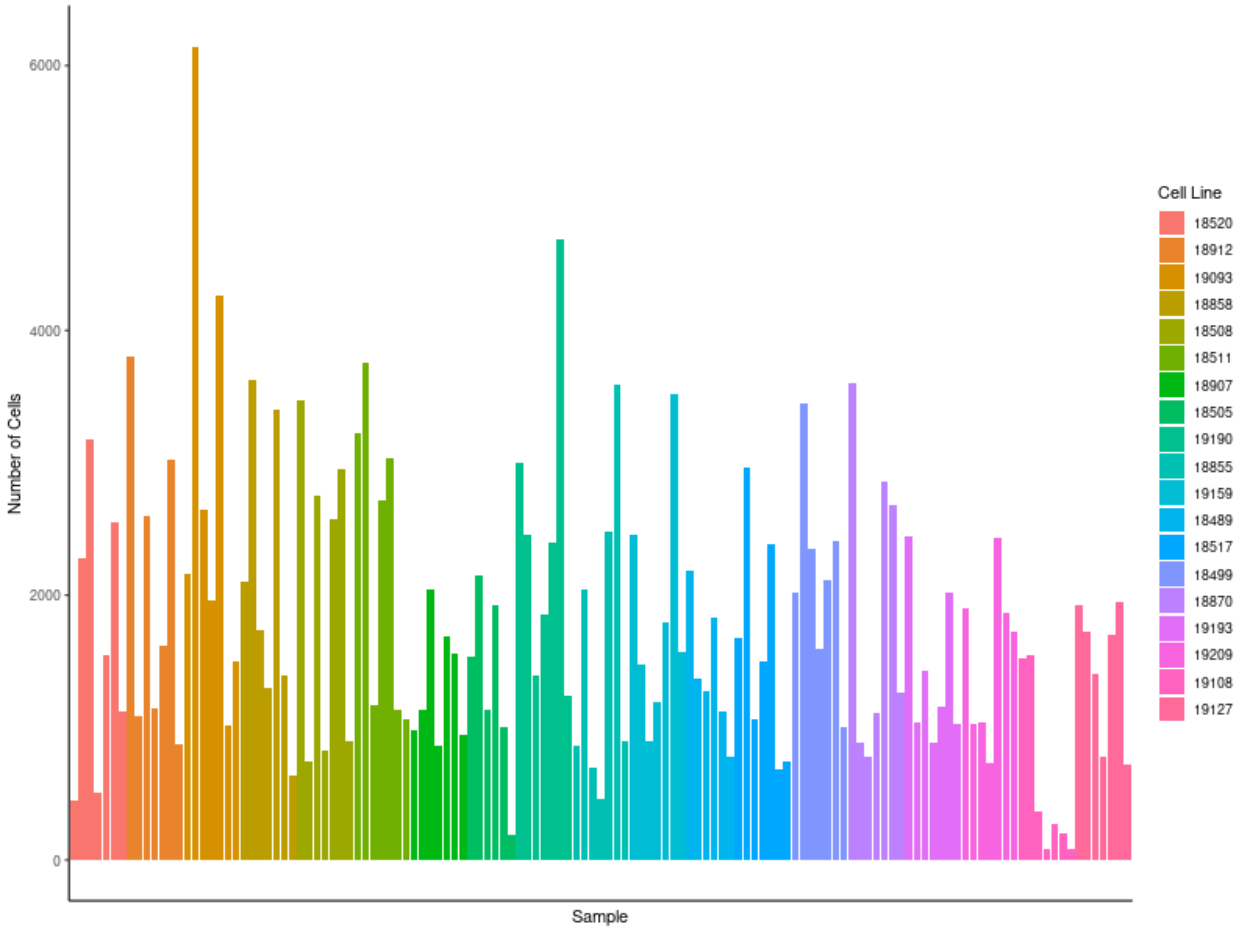


Fig. S14: Number of cells per sample. Number of cells per collected sample following filtering described in Methods (*RNA-seq quantification*).

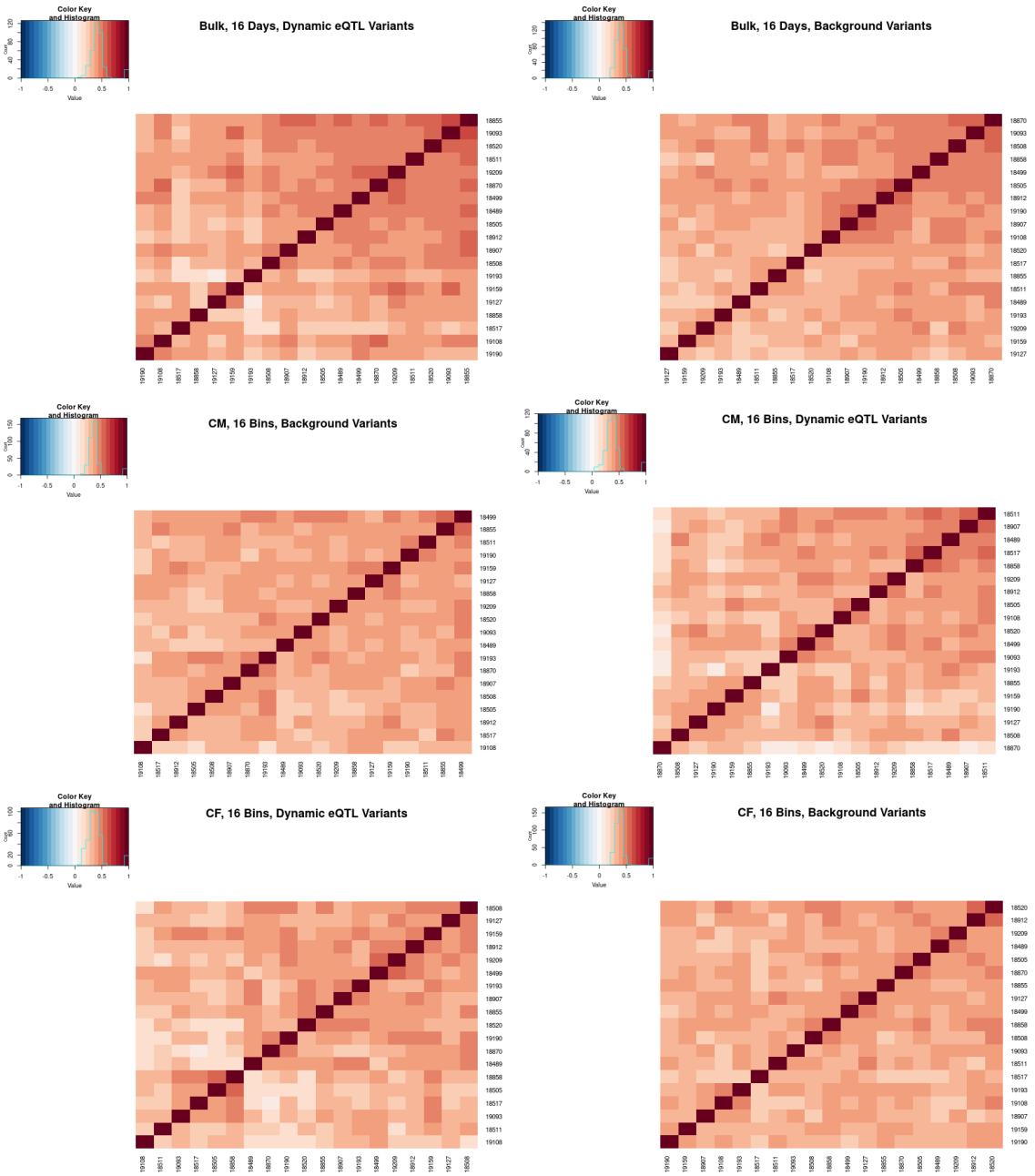


Fig. S15: Genetic correlation across dynamic eQTLs. In order to check whether broad cell line differences are driving false positive dynamic eQTLs, we compared genetic correlation among the top 200 linear dynamic eQTLs for bulk (top, left), and both pseudobulk lineages, cardiomyocyte (middle, left) and cardiac fibroblast (bottom, left), to genetic correlation among a set of background variants within 50kb of a gene, and matched for minor allele frequency (right).

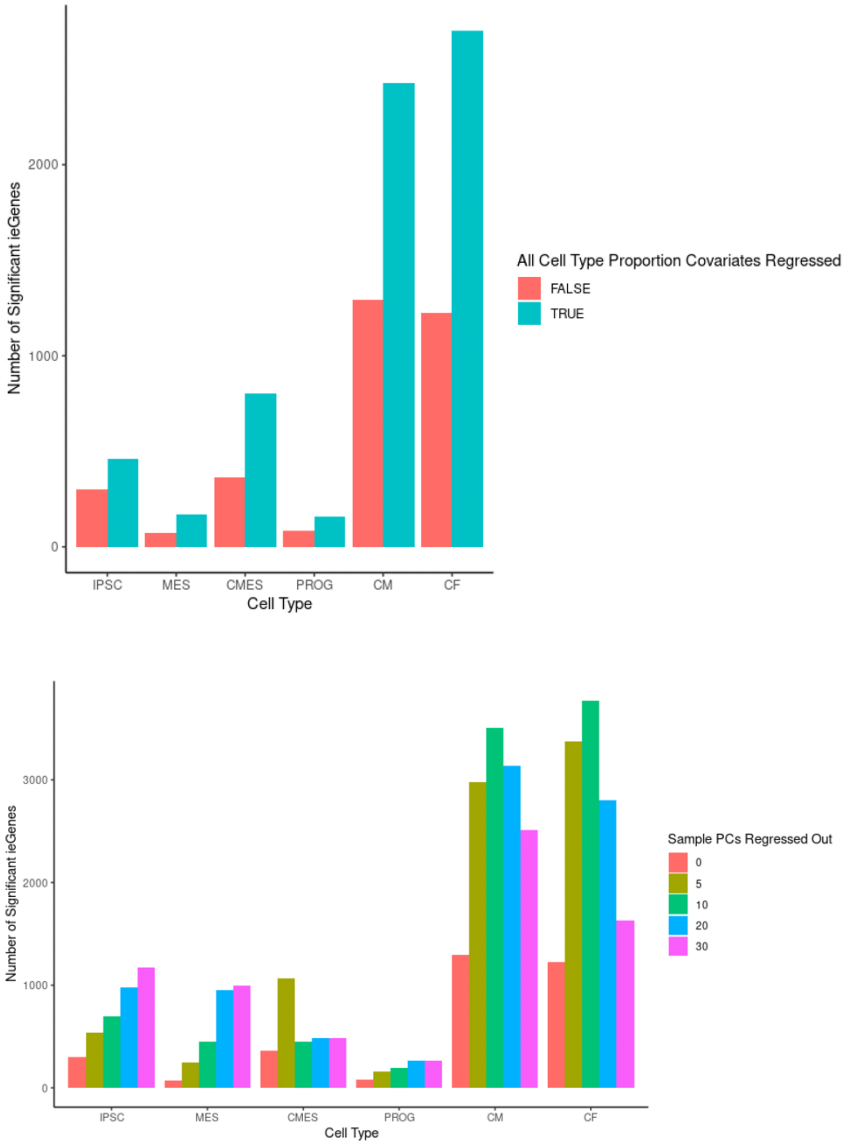


Fig. S16: Impact of additional cell type proportion covariates. We examined the impact of regressing out additional covariates from the interaction eQTL model, and found an increase in the number of genes with a dynamic eQTL, as well as a decrease in the replication rates in bulk dynamic eQTLs (Methods) for both regression of cell type proportions (top) and up to 30 principal components (bottom).

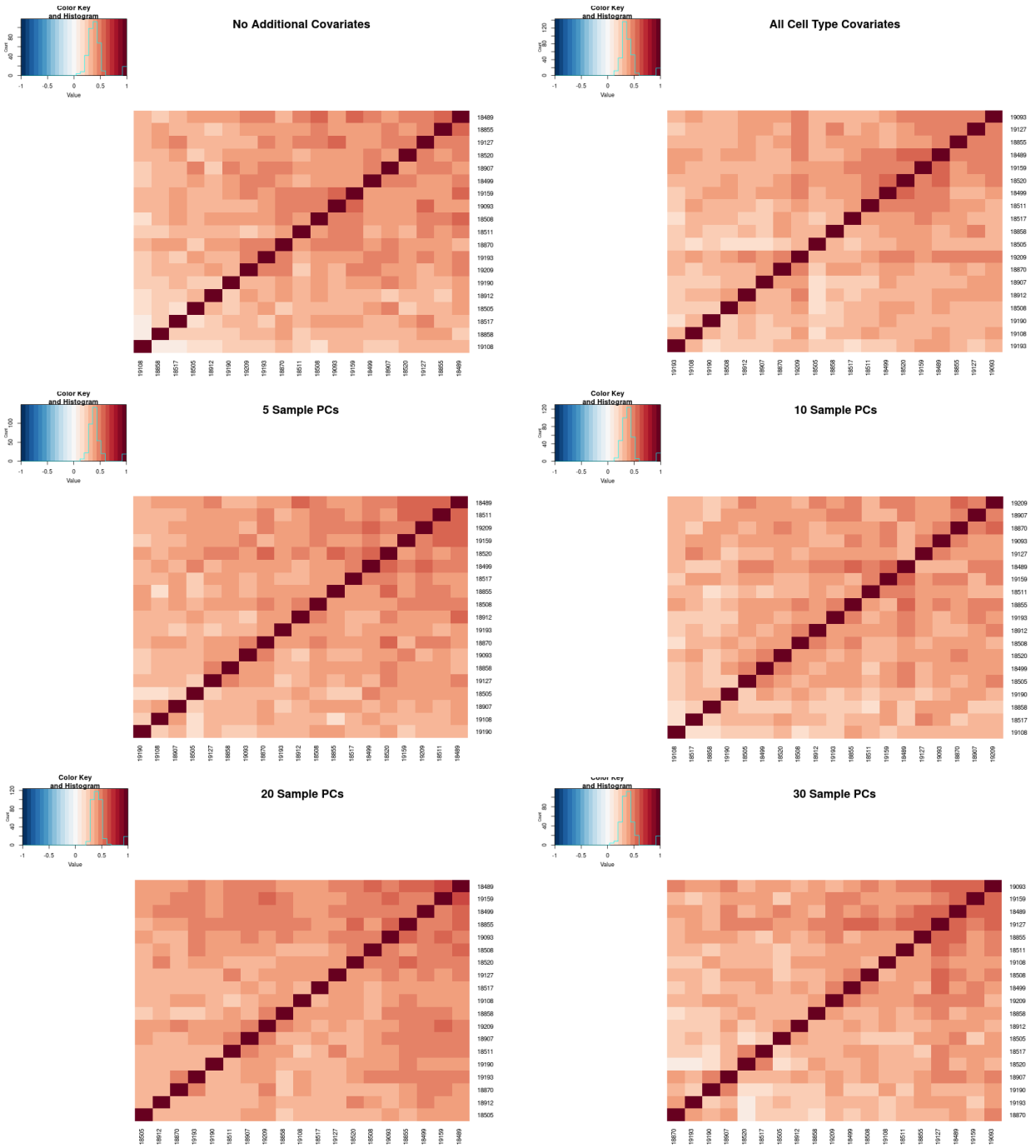


Fig. S17: Genetic correlation across cell type interaction eQTLs. We compared genetic correlation among 200 cardiac fibroblast cell type interaction eQTLs detected exclusively after regressing out additional cell type proportion covariates (a), compared to 200 interaction eQTLs, detected before controlling for cell type proportions (b). We similarly computed genetic correlation among 200 cell type interaction eQTLs discovered only after regression of 5 (c), 10 (d), 20 (e), and 30 (f) sample principal components.

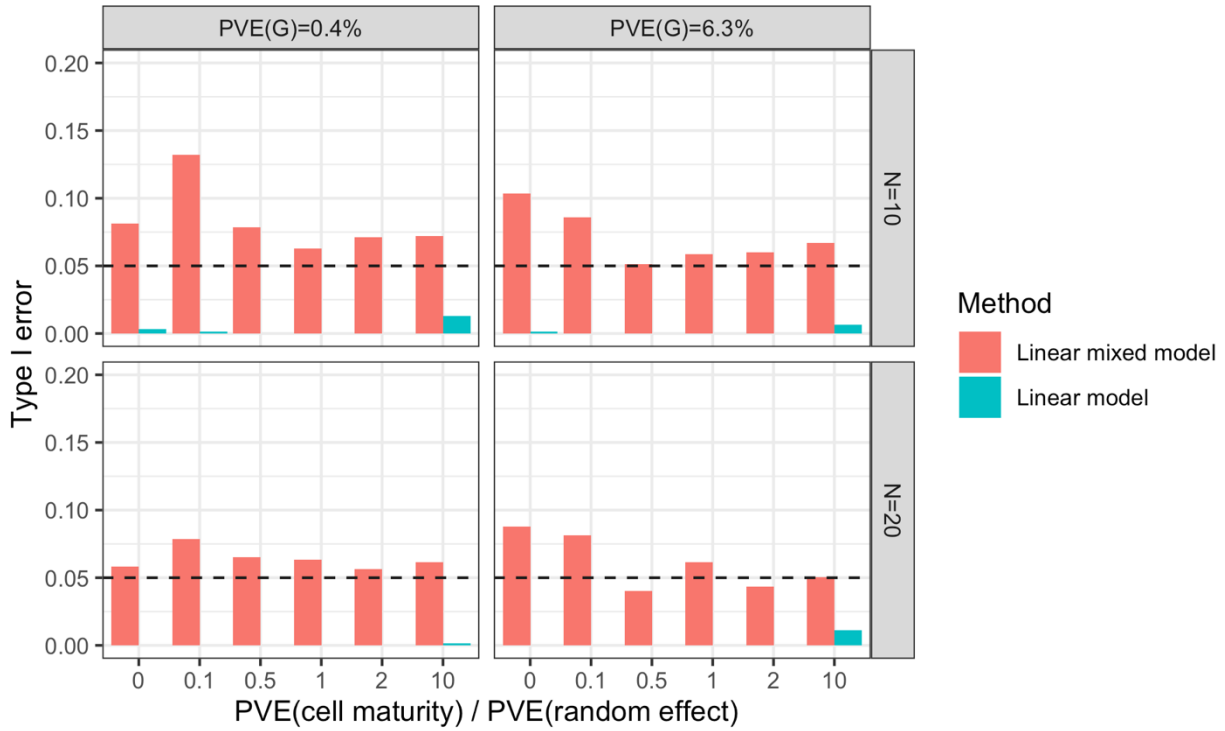


Fig. S18: Selective inference simulations Simulations were performed to examine the impact of selective inference on type I error rates (*Simulations to examine type I errors due to 'double dipping'*). Under the generative model used, inflated type I error rates (bars exceeding the dashed line) were not observed when testing is performed using a linear model (blue).

References (in order of citation):

Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. "Beyond GWASs: Illuminating the Dark Road from Association to Function." *American Journal of Human Genetics* 93 (5): 779–97.

Li, Yang I., Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. 2016. "RNA Splicing Is a Primary Link between Genetic Variation and Disease." *Science* 352 (6285): 600–604.

Albert, Frank W., and Leonid Kruglyak. 2015. "The Role of Regulatory Variation in Complex Traits and Disease." *Nature Reviews. Genetics* 16 (4): 197–212.

The GTEx Consortium. 2020. "The GTEx Consortium atlas of genetic regulatory effects across human tissues." *Science* 369 (6509):1318

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. "Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans." *Nature* 501 (7468): 506–11.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, et al. 2014. "Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals." *Genome Research* 24 (1): 14–24.

Pickrell, Joseph K., John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2010. "Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing." *Nature* 464 (7289): 768–72.

Stranger, Barbara E., Stephen B. Montgomery, Antigone S. Dimas, Leopold Parts, Oliver Stegle, Catherine E. Ingle, Magda Sekowska, et al. 2012. "Patterns of Cis Regulatory Variation in Diverse Human Populations." *PLoS Genetics* 8 (4): e1002639.

Nica, Alexandra C., Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. 2010. "Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations." *PLoS Genetics* 6 (4): e1000895.

Nicolae, Dan L., Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. 2010. "Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS." *PLoS Genetics* 6 (4): e1000888.

Nica, Alexandra C., Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena

1176 Sekowska, Mary Travers et al. 2011. "The architecture of gene regulatory variation
1177 across multiple human tissues: the MuTHER study." *PLoS Genetics* 7 (2): e1002003.
1178
1179 Bis, Joshua C., Maryam Kavousi, Nora Franceschini, Aaron Isaacs, Gonçalo R. Abecasis, Ulf
1180 Schminke, Wendy S. Post, et al. 2011. "Meta-Analysis of Genome-Wide Association
1181 Studies from the CHARGE Consortium Identifies Common Variants Associated with
1182 Carotid Intima Media Thickness and Plaque." *Nature Genetics* 43 (10): 940–47.
1183
1184 Myocardial Infarction Genetics Consortium, Sekar Kathiresan, Benjamin F. Voight, Shaun
1185 Purcell, Kiran Musunuru, Diego Ardissino, Pier M. Mannucci, et al. 2009. "Genome-
1186 Wide Association of Early-Onset Myocardial Infarction with Single Nucleotide
1187 Polymorphisms and Copy Number Variants." *Nature Genetics* 41 (3): 334–41.
1188
1189 Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff,
1190 David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of
1191 Complex Diseases." *Nature* 461 (7265): 747–53.
1192
1193 Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H.
1194 Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the
1195 Underlying Causes of Complex Disease." *Nature Reviews. Genetics* 11 (6): 446–50.
1196
1197 Arvanitis, M., Emmanouil Tampakakis, Yanxiao Zhang, Wei Wang, Adam Auton, 23andMe
1198 Research Team, Diptavo Dutta, Stephanie Glavaris, Ali Keramati, Nilanjan Chatterjee,
1199 Neil C. Chi, Bing Ren, Wendy S. Post & Alexis Battle. 2020. "Genome-wide association
1200 and multi-omic analyses reveal ACTN2 as a gene linked to heart failure." *Nature*
1201 *communications*, 11 (1): 1-12.
1202
1203 Umans, Benjamin D., Alexis Battle, and Yoav Gilad. 2020. "Where Are the Disease-Associated
1204 eQTLs?" *Trends in Genetics: TIG*, September. https://doi.org/10.1016/j.tig.2020.08.009.
1205
1206 Welch, Joshua D., Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and
1207 Evan Z. Macosko. 2019. "Single-cell multi-omic integration compares and contrasts
1208 features of brain cell identity." *Cell*, 177 (7): 1873-1887.
1209
1210 Park, Jihwan, Rojesh Shrestha, Chengxiang Qiu, Ayano Kondo, Shizheng Huang, Max Werth,
1211 Mingyao Li, Jonathan Barasch, and Katalin Suszták. 2018. "Single-cell transcriptomics
1212 of the mouse kidney reveals potential cellular targets of kidney disease." *Science* 360
1213 (6390): 758-763.
1214
1215 Fairfax, Benjamin P., Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen
1216 Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O. Vannberg, and
1217 Julian C. Knight. 2012. "Genetics of gene expression in primary immune cells identifies
1218 cell type–specific master regulators and roles of HLA alleles." *Nature genetics*, *44*(5),

502-510.

Kasela, Silva, Kai Kisand, Liina Tserel, Epp Kaleviste, Anu Remm, Krista Fischer, Tõnu Esko et al. 2017. "Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells." *PLoS genetics*, 13 (3), e1006643.

Kim-Hellmuth, Sarah, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E. Castel et al. 2020. "Cell type–specific genetic regulation of gene expression across human tissues." *Science* 369 (6509).

Strober, B. J., R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, and Y. Gilad. 2019. "Dynamic Genetic Regulation of Gene Expression during Cellular Differentiation." *Science* 364 (6447): 1287–90.

Knowles, David A., Joe R. Davis, Hilary Edgington, Anil Raj, Marie-Julie Favé, Xiaowei Zhu, James B. Potash, et al. 2017. "Allele-Specific Expression Reveals Interactions between Genetic Variation and Environment." *Nature Methods* 14 (7): 699–702.

Taylor, D. Leland, David A. Knowles, Laura J. Scott, Andrea H. Ramirez, Francesco Paolo Casale, Brooke N. Wolford, Li Guan, et al. 2018. "Interactions between Genetic Variation and Cellular Environment in Skeletal Muscle Gene Expression." *PloS One* 13 (4): e0195788.

Fairfax, Benjamin P., Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, et al. 2014. "Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression." *Science* 343 (6175): 1246949.

Smirnov, Denis A., Michael Morley, Eunice Shin, Richard S. Spielman, and Vivian G. Cheung. 2009. "Genetic Analysis of Radiation-Induced Changes in Human Gene Expression." *Nature* 459 (7246): 587–91.

Watts, Jason A., Michael Morley, Joshua T. Burdick, Jennifer L. Fiori, Warren J. Ewens, Richard S. Spielman, and Vivian G. Cheung. 2002. "Gene Expression Phenotype in Heterozygous Carriers of Ataxia Telangiectasia." *American Journal of Human Genetics* 71 (4): 791–800.

Kariuki, Silvia N., Joseph C. Maranville, Shaneen S. Baxter, Choongwon Jeong, Shigeki Nakagome, Cara L. Hrusch, David B. Witonsky, Anne I. Sperling, and Anna Di Rienzo. 2016. "Mapping Variation in Cellular and Transcriptional Response to 1,25-Dihydroxyvitamin D3 in Peripheral Blood Mononuclear Cells." *PloS One* 11 (7): e0159779.

Alleyne, Dereck, David B. Witonsky, Brandon Mapes, Shigeki Nakagome, Meredith Sommars, Ellie Hong, Katy A. Muckala, Anna Di Rienzo, and Sonia S. Kupfer. 2017. "Colonic Transcriptional Response to 1α,25(OH) Vitamin D in African- and European-

Americans." *The Journal of Steroid Biochemistry and Molecular Biology* 168 (April): 49–59.

Pijuan-Sala, Blanca, Carolina Guibentif, and Berthold Göttgens. 2018. "Single-Cell Transcriptional Profiling: A Window into Embryonic Cell-Type Specification." *Nature Reviews. Molecular Cell Biology* 19 (6): 399–412.

Cuomo, Anna S. E., Daniel D. Seaton, Davis J. McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, et al. 2020. "Single-Cell RNA-Sequencing of Differentiating iPS Cells Reveals Dynamic Genetic Effects on Gene Expression." *Nature Communications* 11 (1): 810.

Jerber, Julie, Daniel D. Seaton, Anna SE Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel et al. 2021. "Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation." *Nature genetics* 53 (3): 304-312.

Westra, Harm-Jan, Danny Arends, Tõnu Esko, Marjolein J. Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen et al. 2015. "Cell specific eQTL analysis without sorting cells." *PLoS genetics*, 11 (5): e1005223.

Selewa, Alan, Ryan Dohn, Heather Eckart, Stephanie Lozano, Bingqing Xie, Eric Gauchat, Reem Elorbany, et al. 2020. "Systematic Comparison of High-Throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation." *Scientific Reports* 10 (1): 1535.

Burridge, Paul W., Elena Matsa, Praveen Shukla, Ziliang C. Lin, Jared M. Churko, Antje D. Ebert, Feng Lan et al. 2014. "Chemically defined generation of human cardiomyocytes." *Nature methods*, 11 (8): 855-860.

Ahmad, Ferhaan, Sanjay K. Banerjee, Michele L. Lage, Xueyin N. Huang, Stephen H. Smith, Samir Saba, Jennifer Rager, et al. 2008. "The Role of Cardiac Troponin T Quantity and Function in Cardiac Development and Dilated Cardiomyopathy." *PloS One* 3 (7): e2642.

Bizy, Alexandra, Guadalupe Guerrero-Serna, Bin Hu, Daniela Ponce-Balbuena, B. Cicero Willis, Manuel Zarzoso, Rafael J. Ramirez, et al. 2013. "Myosin Light Chain 2-Based Selection of Human iPSC-Derived Early Ventricular Cardiac Myocytes." *Stem Cell Research* 11 (3): 1335–47.

Ieda, Masaki, Takatoshi Tsuchihashi, Kathryn N. Ivey, Robert S. Ross, Ting-Ting Hong, Robin M. Shaw, and Deepak Srivastava. 2009. "Cardiac Fibroblasts Regulate Myocardial Proliferation through beta1 Integrin Signaling." *Developmental Cell* 16 (2): 233–44.

Zhang, Jianhua, Ran Tao, Katherine F. Campbell, Juliana L. Carvalho, Edward C. Ruiz, Gina C. Kim, Eric G. Schmuck, et al. 2019. "Functional Cardiac Fibroblasts Derived from Human Pluripotent Stem Cells via Second Heart Field Progenitors." *Nature Communications* 10 (1): 2238.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. "SCANPY: large-scale single-cell gene expression data analysis." *Genome biology*, 19 (1): 1-5.

Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software." *PloS one*, 9 (6), e98679.

Haghverdi, Laleh, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. "Diffusion pseudotime robustly reconstructs lineage branching." 2016. *Nature methods*, *13*(10), 845.

Wolf, F. Alexander, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. 2019. "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells." *Genome biology* 20 (1): 1-9

Barbeira, Alvaro N., Rodrigo Bonazzola, Eric R. Gamazon, Yanyu Liang, YoSon Park, Sarah Kim-Hellmuth, Gao Wang et al. 2021. "Exploiting the GTEx resources to decipher the mechanisms at GWAS loci." *Genome biology*, 22 (1): 1-24.

Loirand, Gervaise, and Pierre Pacaud. 2014. "Involvement of Rho GTPases and their regulators in the pathogenesis of hypertension." *Small GTPases*, 5 (4): e983866.

Newman, Aaron M., Chloé B. Steen, Chih Long Liu, Andrew J. Gentles, Aadel A. Chaudhuri, Florian Scherer, Michael S. Khodadoust et al. 2019. "Determining cell type abundance and expression from bulk tissues with digital cytometry." *Nature biotechnology*, 37 (7): 773-782.

Nakagami, Hironori, Yasushi Kikuchi, Tomohiro Katsuya, Ryuichi Morishita, Hiroshi Akasaka, Shigeyuki Saitoh, Hiromi Rakugi, Yasufumi Kaneda, Kazuaki Shimamoto, and Toshio Ogihara. 2007. "Gene polymorphism of myospryn (cardiomyopathy-associated 5) is associated with left ventricular wall thickness in patients with hypertension." *Hypertension research*, 30 (12): 1239-1246.

D'Antonio-Chronowska, Agnieszka, Margaret K. R. Donovan, William W. Young Greenwald, Jennifer Phuong Nguyen, Kyohei Fujita, Sherin Hashem, Hiroko Matsui, Francesca Soncin, Mana Parast, Michelle C. Ward, Florence Coulet, Erin N. Smith, Eric Adler, Matteo D'Antonio, and Kelly A. Frazer. 2019. "Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories." *Stem Cell Reports* 13 (5): 924–38.

Brade, Thomas, Luna S. Pane, Alessandra Moretti, Kenneth R. Chien, and Karl-Ludwig Laugwitz. 2013. “Embryonic Heart Progenitors and Cardiogenesis.” *Cold Spring Harbor Perspectives in Medicine* 3 (10): a013847.

Taylor, Jonathan, and Robert J. Tibshirani. 2015. "Statistical learning and selective inference." *Proceedings of the National Academy of Sciences*, *112*(25), 7629-7634.

Gao, Lucy L., Jacob Bien, and Daniela Witten. 2020. "Selective Inference for Hierarchical Clustering." *arXiv preprint* arXiv:2012.02936.

Chung, Neo Christopher, and John D. Storey. 2015. "Statistical significance of variables driving systematic variation in high-dimensional data." *Bioinformatics*, 31 (4): 545-554.

Barrett, Tanya, Tugba O. Suzek, Dennis B. Troup, Stephen E. Wilhite, Wing-Chi Ngau, Pierre Ledoux, Dmitry Rudnev, Alex E. Lash, Wataru Fujibuchi, and Ron Edgar. 2005. "NCBI GEO: mining millions of expression profiles—database and tools." *Nucleic acids research*, 33 (suppl 1): D562-D566.

(Methods References)

Banovich, Nicholas E., Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, et al. 2018. “Impact of Regulatory Variation across Human iPSCs and Differentiated Cells.” *Genome Research* 28 (1): 122–31.

Lian, Xiaojun, Jianhua Zhang, Samira M. Azarin, Kexian Zhu, Laurie B. Hazeltine, Xiaoping Bao, Cheston Hsiao, Timothy J. Kamp, and Sean P. Palecek. 2013. “Directed Cardiomyocyte Differentiation from Human Pluripotent Stem Cells by Modulating Wnt/β-Catenin Signaling under Fully Defined Conditions.” *Nature Protocols* 8 (1): 162–75.

Tohyama, Shugo, Fumiyuki Hattori, Motoaki Sano, Takako Hishiki, Yoshiko Nagahata, Tomomi Matsuura, Hisayuki Hashimoto, et al. 2013. “Distinct Metabolic Flow Enables Large-Scale Purification of Mouse and Human Pluripotent Stem Cell-Derived Cardiomyocytes.” *Cell Stem Cell* 12 (1): 127–37.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.

Smith, Tom, Andreas Heger, and Ian Sudbery. 2017. “UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy.” *Genome Research* 27 (3): 491–99.

Kang, Hyun Min, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, et al. 2018. “Multiplexed Droplet Single-Cell RNA-Sequencing Using Natural Genetic Variation.” *Nature Biotechnology* 36 (1): 89–94.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "Comprehensive integration of single-cell data." *Cell* 177 (7): 1888-1902.

Hafemeister, Christoph, and Rahul Satija. 2019. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression." *Genome biology*, 20 (1): 1-15.

Cakir, Batuhan, Martin Prete, Ni Huang, Stijn Van Dongen, Pinar Pir, and Vladimir Yu Kiselev. 2020. "Comparison of visualization tools for single-cell RNAseq data." *NAR Genomics and Bioinformatics*, *2*(3), lqaa052.

Robinson, Mark D., and Alicia Oshlack. 2010. "A scaling normalization method for differential expression analysis of RNA-seq data." *Genome biology*, 11 (3): 1-9.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, 26 (1): 139-140.

Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. “DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation.” *Nature* 482 (7385): 390–94.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge et al. 2019. "GENCODE reference annotation for the human and mouse genomes." *Nucleic acids research*, 47 (D1): D766-D773.

Storey, John D. 2003. "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics*, 31 (6): 2013-2035.