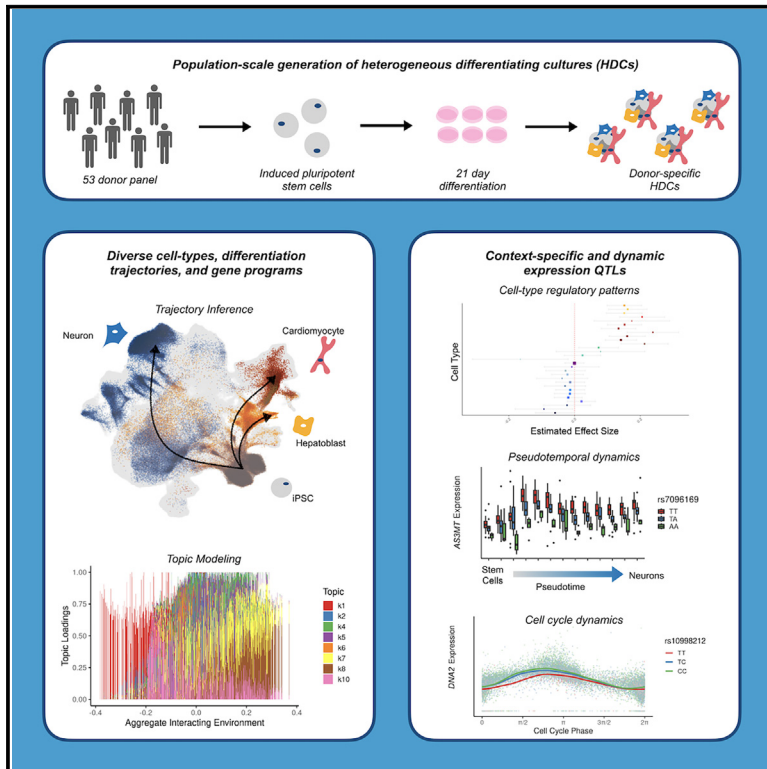# Cell type and dynamic state govern genetic regulation of gene expression in heterogeneous differentiating cultures

Graphical abstract

Highlights

- Dozens of cell types in heterogeneous differentiating cultures from 53 human cell lines

- Dynamic genetic regulation of gene expression in diverse differentiation trajectories

- Previously inaccessible contexts reveal functional impact of disease loci

Authors

Joshua M. Popp, Katherine Rhodes, Radhika Jangi, ..., Karl Tayeb, Alexis Battle, Yoav Gilad

Correspondence

ajbattle@jhu.edu (A.B.),
gilad@uchicago.edu (Y.G.)

In brief

Popp et al. generate dozens of cell types from 53 human iPSC lines in order to characterize the dynamic genetic regulation of gene expression across early stages of cellular differentiation. Accessing these understudied contexts can clarify the functional impact of disease loci with unknown mechanisms of action.

Cell Genomics

Article

Cell type and dynamic state govern genetic regulation of gene expression in heterogeneous differentiating cultures

Joshua M. Popp,[1,8] Katherine Rhodes,[2,8] Radhika Jangi,[3] Mingyuan Li,[3] Kenneth Barr,[2] Karl Tayeb,[4] Alexis Battle,[1,5,6,*] and Yoav Gilad[2,7,9,*]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[3]Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA
[4]Committee on Genetics, Genomics, and Systems Biology, University of Chicago, Chicago, IL 60637, USA
[5]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[6]Department of Genetic Medicine, Johns Hopkins University, Baltimore, MD 21218, USA
[7]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
[8]These authors contributed equally
[9]Lead contact
*Correspondence: ajbattle@jhu.edu (A.B.), gilad@uchicago.edu (Y.G.)
https://doi.org/10.1016/j.xgen.2024.100701

SUMMARY

Identifying the molecular effects of human genetic variation across cellular contexts is crucial for understanding the mechanisms underlying disease-associated loci, yet many cell types and developmental stages remain underexplored. Here, we harnessed the potential of heterogeneous differentiating cultures (HDCs), an *in vitro* system in which pluripotent cells asynchronously differentiate into a broad spectrum of cell types. We generated HDCs for 53 human donors and collected single-cell RNA sequencing data from over 900,000 cells. We identified expression quantitative trait loci in 29 cell types and characterized regulatory dynamics across diverse differentiation trajectories. This revealed novel regulatory variants for genes involved in key developmental and disease-related processes while replicating known effects from primary tissues and dynamic regulatory effects associated with a range of complex traits.

INTRODUCTION

Decoding the molecular consequences of genetic variation is a central goal in human genetics. With the advent of genome-wide association studies (GWASs), a vast array of genetic variants associated with diseases have been uncovered. These predominantly lie in non-coding regions of the genome, suggesting primarily regulatory mechanisms.[1] This insight has spurred a surge in mapping expression quantitative trait loci (eQTLs) to understand how the disease-associated genetic variants influence gene expression levels. Despite significant strides made by several large-scale projects, such as the GTEx Consortium, to map eQTLs,[2–7] a comprehensive understanding of the molecular impacts of disease-associated loci remains elusive, in part due to the context-dependent and dynamic nature of gene regulation.[8–11]

Gene regulation varies by contexts including cell type, temporal stage, and environment. This poses a formidable challenge for human studies that seek to characterize the gene regulatory basis for complex traits.[12–14] Studies using postmortem human tissues, while delivering important insight, often fall short of capturing the full spectrum of dynamic regulatory effects because they reflect predefined adult tissue contexts and

because most studies have utilized bulk sequencing. Recent advances in single-cell technologies have started to shift this paradigm by enabling researchers to collect a heterogeneous biological sample and disentangle context-specific regulatory variation through downstream analysis of single-cell molecular phenotypes.[15–20] Still, many contexts are difficult to sample from healthy human donors, particularly dynamic contexts where we would like to capture multiple time points from the same individual. This would be nearly impossible for an inaccessible tissue. This has motivated the use of differentiation protocols for *in vitro* cell cultures, which have offered access to dynamic regulatory effects, including fleeting effects present only at intermediate stages of differentiation.[21] These *in vitro* systems are each imperfect reflections of human cell biology, but the activation of a range of relevant *cis*-regulatory elements can reveal the effects of variants within them even without completely recapitulating the *in vivo* cellular state. Indeed, studies in these systems have captured regulatory effects of numerous disease-associated loci and variants near genes involved in developmental processes.[15,16,21] However, most protocols would require separate experimental setups for each cell type or perturbation of interest, making it difficult to efficiently explore the space of disease-relevant contexts.

In this study, we explored gene regulation across diverse cellular contexts using heterogeneous differentiating cultures (HDCs), terminology we introduce as a broad descriptor encompassing a class of related *in vitro* models that can be used to explore diverse cellular contexts efficiently. Here, we focus on unguided HDCs, which are based on embryoid body systems using an extended culturing time to consistently generate dozens of cell types derived from all three developmental germ layers.[22,23] We have also developed guided HDCs, which push induced pluripotent stem cells (iPSCs) toward certain lineages to enrich for multiple cell types within a particular tissue and offer the ability to pursue more targeted questions. While both unguided and guided HDCs differ from *in vivo* cellular biology as expected, we have demonstrated in previous studies and here that the expression profiles and genetics effects found in HDCs overlap those found using primary cell types and tissues.[24,25]

Here, we generated unguided HDCs from a panel of 53 human iPSC lines and measured gene expression at single-cell resolution in over 900,000 cells. We mapped eQTLs in 29 cell types, including many that have never before been characterized at the population level in humans, and identified dynamic genetic effects on gene regulation that vary with respect to diverse differentiation trajectories and gene programs.

RESULTS

HDCs generate diverse cell types

We established unguided HDCs from the iPSCs of 53 unrelated Yoruba individuals from Ibadan, Nigeria (YRI) (Figure 1A; STAR Methods). Briefly, we formed HDCs in batches of 4–8 individuals and maintained them in culture for 21 days (Table S1). Within each batch, we formed, maintained, and dissociated HDCs in parallel. After dissociation, we multiplexed samples from each individual in equal proportions in preparation for single-cell RNA sequencing (RNA-seq), targeting a depth of 100,000 reads per cell. After quality control, filtering, and de-multiplexing, we retained data from 909,536 high-quality cells (median: 1,241 cells per individual per replicate; median: 21,990 unique molecular identifier [UMI] counts per cell).

To initially assess cellular diversity in HDCs, we developed a cell type classifier based on data and annotations from the fetal cell atlas.[26] We used a curated set of 33 high-confidence cell type labels (STAR Methods; Figure S1) that span the three main germ layers (Figure 1B). We assigned 651,129 cells (72% of all cells) to one of these cell types based on gene expression signatures (Table S2). Using this approach, 28% of the cells remained unclassified. Of the 33 cell types, 29 are represented with a minimum of 5 cells in at least 25 individuals. While the proportion of cells of each type varied between individuals (Figure 1C), 52 of 53 individuals have data from at least 5 cells from most cell types (median: 29 of 33 cell types per donor; Figure 1D). Some of the unannotated cells express markers of pluripotency, suggesting that asynchronous differentiation within HDCs enabled us to collect pluripotent cells alongside partially and fully differentiated cell types. Indeed, manually adding a marker gene set for iPSCs[27] to the list of cell type signatures enabled us to classify 21,370 previously unannotated cells as iPSCs (Figure 1B). Since this iPSC signa-

ture was obtained from a separate reference, and since eQTLs in iPSCs have previously been thoroughly characterized,[28,29] we focused on the 29 common fetal cell atlas cell types in the subsequent cell-type-stratified eQTL analysis, filtering to the 52 donors that successfully differentiated into diverse cell types (Table S3). We re-incorporate these pluripotent and un-annotated cells in later analyses that focused on evaluating regulatory dynamics across the HDC system.

eQTLs across cell types

We mapped genetic effects on gene regulation in each of the 29 discretely defined fetal cell atlas cell types. To mitigate the effects of noise inherent to single-cell data, we aggregated single-cell expression into pseudobulk such that each observation represents all individual cells from a single donor/cell type combination[30,31] (STAR Methods). This aggregation step also allowed us to take advantage of well-established methods for eQTL mapping using bulk RNA sequencing data.

We initially performed *cis* eQTL mapping specifically in each cell type, limiting our analysis to the 29 annotated cell types with at least 5 cells from at least 25 donors (Figure 2A). We included expression principal components for data from each cell type as covariates to control for hidden factors driving global expression variability, including batch effects.[32] Across all cell types, we identified a total of 31,179 eQTLs (associated with 2,114 eGenes) at a global q value cutoff of 0.05.[33] 79% of these HDC eQTLs were previously identified by GTEx (Figure 2B); that is, 6,572 of our eQTLs have not been identified in healthy adult tissues. The HDCs include many developing cell types not found in GTEx adult tissues. Indeed, the subset of eGenes regulated by eQTLs identified in HDCs, but not in GTEx, were enriched for several developmental processes, including tissue development (odds ratio [OR] = 2.08, two-sided Fisher's exact test p = 4.5e−5), central nervous system development (OR = 2.31, p = 1.2e−4), and circulatory system development (OR = 2.32, p = 1.3e−4) (Table S4). The clustering of non-GTEx HDC eQTLs upstream of transcription start sites (Figure S2) indicates that these enrichments are not an artifact due to false positive associations near developmental genes that are relatively depleted for eQTLs in GTEx.[11] A subset of these non-GTEx HDC eQTLs (2,705 of 6,572) have been previously characterized in iPSCs.[6] The HDC system offers access to eQTLs that replicate in diverse primary adult tissues in an *in vitro* setting amenable to environmental and genomic perturbation, at a breadth that has not been feasible in existing studies of eQTLs in iPSCs[6] or iPSC-derived cell types.[15–17,28,34]

Within each cell type, we identified a median of 2,099 eQTLs (maximum of 15,064 eQTLs in peripheral nervous system [PNS] glial cells) in a median of 126 eGenes (maximum of 919 eGenes in PNS glial cells) (Table S5). The number of eQTLs detected in each cell type is correlated with the median number of individual cells from which we have data per cell type across individuals (Figure S3), suggesting that the power to detect eQTLs is limited by both sample size and the number of captured individual cells. To account for incomplete power to detect eQTLs in any given cell type and to assess the extent of eQTL sharing between cell types, we analyzed the data using multivariate adaptive shrinkage (mash).[35] By borrowing information
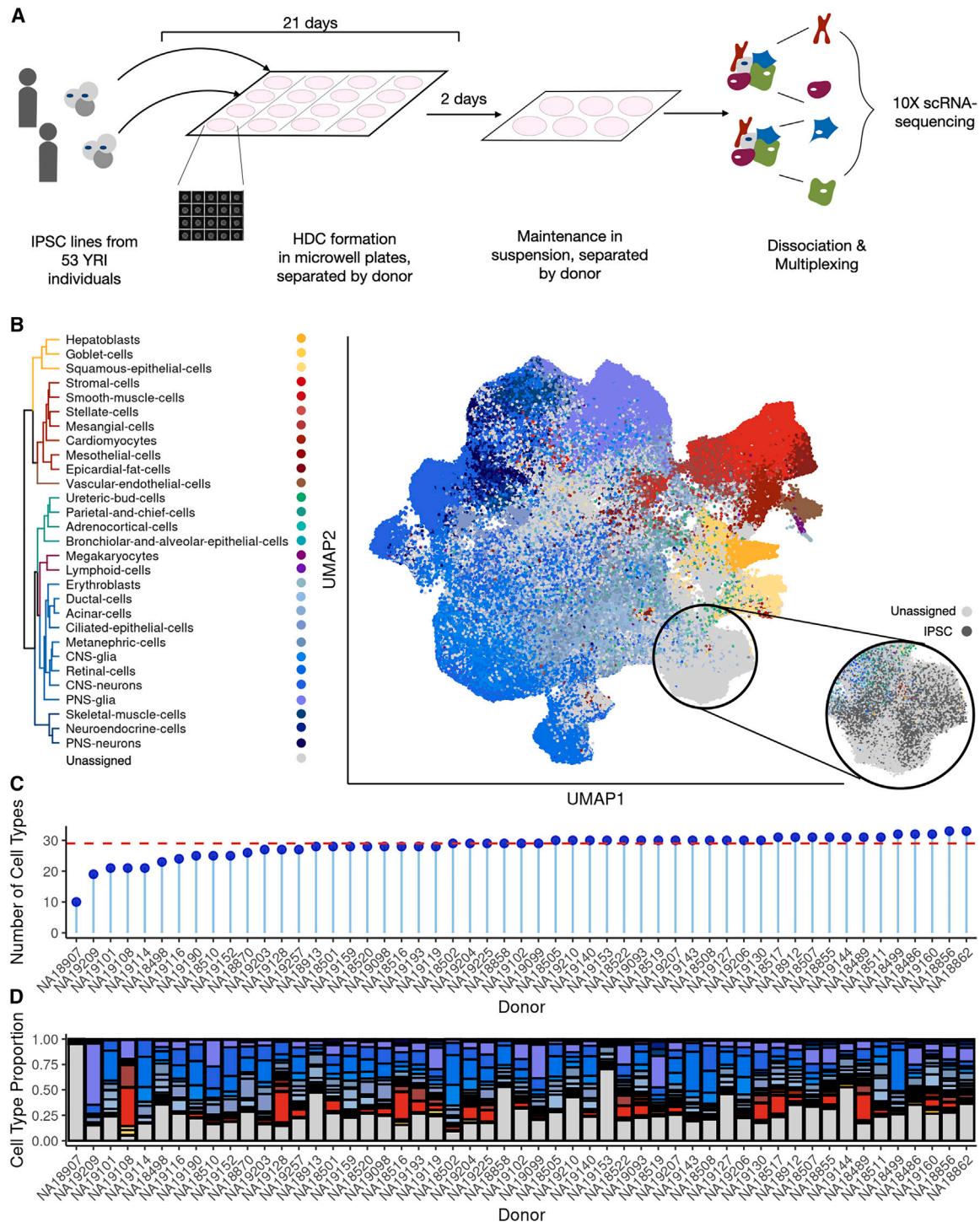
Figure 1. HDC panel from 53 human iPSC lines

(A) Schematic illustration of the data generation process.

(B) Uniform manifold approximation and projection (UMAP) embedding of over 900,000 heterogeneous differentiating culture (HDC) cells annotated using the fetal cell atlas. Inset shows annotation of a primarily unassigned group of cells with an augmented classifier trained with pluripotent cells.

(C) Most cell lines generate most cell types with sufficient coverage for inclusion in QTL analyses.

(D) Cell type proportions vary between lines.
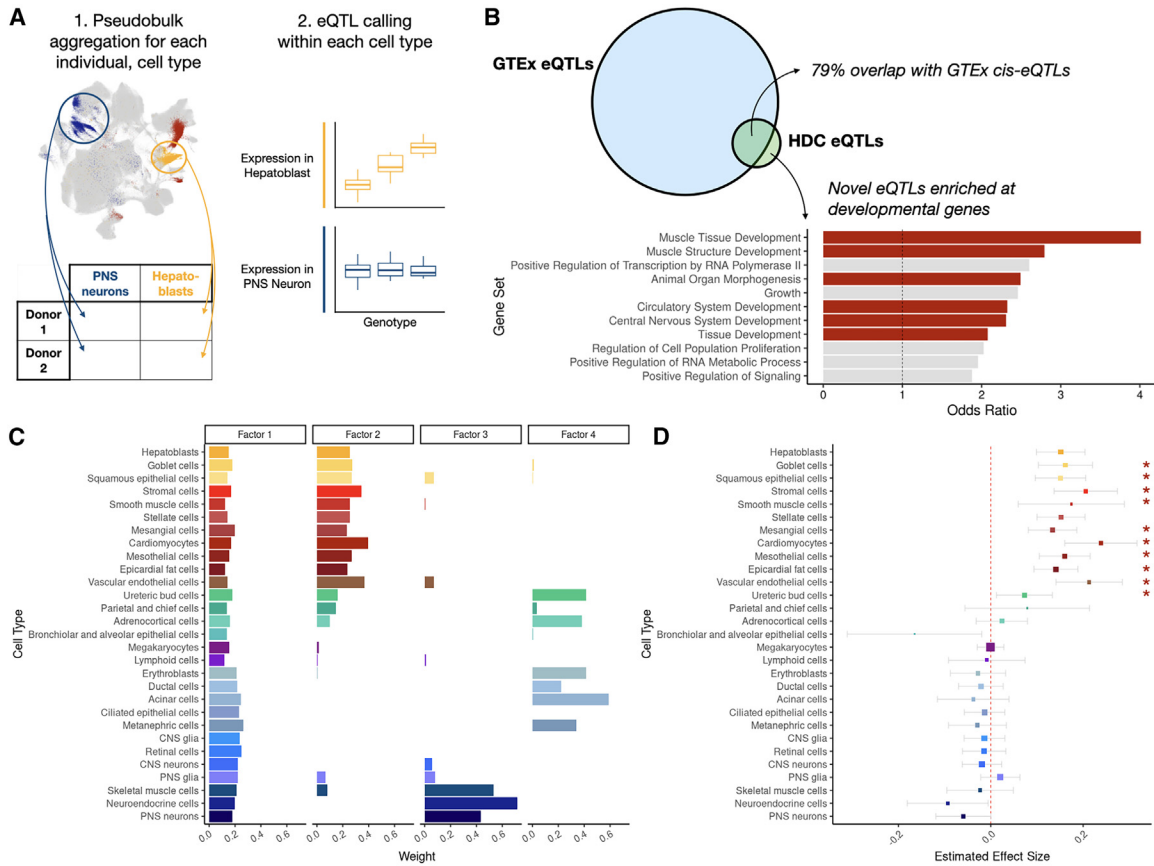
Figure 2. eQTL calling across 29 cell types

(A) To perform eQTL calling, UMI counts from cells from the same donor and cell type were aggregated into a pseudobulk sample, and eQTL calling was performed separately in each cell type.

(B) Comparison of HDC eQTLs to GTEx eQTLs. Bar plot shows odds ratio of 11 Gene Ontology (GO) biological processes significantly enriched for HDC eGenes with no eQTLs overlapping GTEx hits (FDR ≤ 0.05, background gene set all HDC eGenes). Developmental processes are highlighted in red.

(C) Patterns of regulatory effects learned through matrix decomposition of eQTL effect sizes in each cell type. Beyond the densely loaded factor 1, remaining factors partition cell types of similar developmental origins.

(D) Metaplot of eQTL for the gene *SH3PXD2B* at rs10042482. Boxes are centered at the estimated posterior mean effect in the given cell type. Error bars show ±posterior standard deviation, and box size indicates precision (1/squared posterior standard deviation). This eQTL was not detected in GTEx or the univariate (cell type-by-cell type) analysis. * indicates significance at a local false sign rate of 0.05.

across cell types and genes, mash allows us to detect weak signals that emerge repeatedly in different cell types with greater confidence, thereby improving power. More generally, mash can be used to identify major patterns of heterogeneity and sharing between cell types.

The four strongest regulatory patterns revealed when we used this approach depict broadly shared regulation among all cell types and then among subsets of developmentally related cells. Of these more specific patterns, the first corresponds to endoderm- and mesoderm-derived cell types, the second to ectoderm-derived cell types, and the third to the cluster of epithelial cell types, which may correspond to the neuroepithelium, which would not be expected to appear in our fetal cell atlas reference (Figure 2C). We incorporated these candidate regulatory patterns alongside several "canonical" patterns such as single cell type effects into a prior distribution that guides hypothesis testing (STAR Methods).

Using mash, we detected an additional 56,614 eQTLs corresponding to 4,973 eGenes (Table S6). The example of *SH3PXD2B* (Figure 2D), a gene thought to be important for cardiac and skeletal development,[36] highlights the utility of considering structured regulatory patterns in the analysis of heterogeneous single-cell data. Indeed, no significant eQTLs for *SH3PXDB2* were found in GTEx, nor the univariate analysis considering one HDC cell type at a time (Figure S4). Specifically, when we leveraged regulatory patterns shared by cell types of similar developmental origin, we found a significant *cis* eQTL for *SH3PXDB2* (local false sign rate [lfsr] < 0.05) in nearly all endoderm- and mesoderm-derived cell types.

Dynamic genetic regulation along diverse differentiation trajectories

Next, we took a different approach to characterize how interactions between genetic variants and the cellular environment

impact gene expression levels. To do so, we abandon discretely defined cell type labels in favor of more continuous and nuanced representations of cellular variation.

We were particularly interested in characterizing temporally dynamic genetic effects, including effects that fluctuate during cellular development. As cells within HDCs differentiate asynchronously, our current dataset captures continuous gene regulatory variation along multiple developmental trajectories. To validate our approach for defining differentiation trajectories in HDCs, we began by focusing on the cardiomyocyte lineage because we had previously collected single-cell data from directly differentiated cardiomyocytes using a time course study design, which can be used as the ground truth for both gene expression and regulatory dynamics. First, we compiled manually curated gene lists containing marker genes from each stage of cardiomyocyte differentiation (STAR Methods; Table S7).[17] We used a cell scoring tool (single-cell disease relevance scoring [scDRS])[37] to identify subpopulations of cells with enriched expression for gene sets specific to the cardiomyocyte trajectory and applied principal-component analysis to infer the pseudotime (STAR Methods; Figure 3A).[16,38] The first expression principal component using data from these cardiomyocyte trajectory cells offers a reasonable pseudotime metric (Figure 3B), as it captures sequential expression of marker genes for each stage of cardiomyocyte differentiation (*NANOG*, an iPSC marker gene; *MIXL1*, mesendoderm; *MESP1*, mesoderm; *GATA4*, cardiac progenitor; and *TNNT2*, cardiomyocyte).[39] Following our previous approach,[21] we identified 709 linear dynamic eQTLs (in 47 eGenes) along the cardiomyocyte trajectory with a less stringent genome-wide false discovery rate (FDR) cutoff of 0.1, using EigenMT to control for multiple testing burden within each gene (STAR Methods). The majority of these effects were replicated in the directly differentiated cardiomyocytes, where the sample size was far smaller ($n = 19$; π_1 replication rate = 0.58).[17]

After validating our approach in the cardiomyocyte trajectory, which is derived from mesoderm, we examined neuronal and hepatic differentiation trajectories, which, respectively, represent the ectodermal and endodermal germ layers (Tables S8 and S9). At a genome-wide FDR of 0.1, we found 1,965 dynamic eQTLs (166 eGenes) along the neuronal differentiation trajectory and 472 dynamic eQTLs (44 eGenes) along the hepatocyte differentiation trajectory (Table S10).

We further categorized the dynamic eQTLs we identified in these three developmental trajectories into early (40%), late (57%), and switch (3%) effects, excluding three genes where eQTL classification diverged between trajectories (see STAR Methods). For example, *RILPL1*, which is thought to regulate cell shape and polarity,[40] is associated with an early dynamic eQTL in all three trajectories (Figure 3C), while *ACAA2*, which encodes a protein involved in fatty acid metabolism in the liver, is associated with a late eQTL only in hepatocytes (Figure 3D). While 77% of late and switch dynamic eQTLs overlap previously discovered effects listed in the GTEx catalog, similar to the proportion of overlapping cell type eQTLs, the overlap with GTEx drops to 60% for early dynamic eQTLs. That is, eQTLs identified only in the early-developing cell types of each trajectory are markedly less likely to be observed in GTEx adult tissue eQTLs.

Resolving complex regulatory interactions using topic analysis

Cell type and differentiation stage represent the most salient aspects of cellular identity in our dataset. Regulatory changes along differentiation and between cell types lead to strong gene expression differences, such that the straightforward application of unsupervised machine learning methods (such as clustering and principal-component analysis) to gene expression will first stratify cells based on these features. However, as they differentiate, cells are simultaneously engaging in a wide array of dynamic processes, including growth, division, and signaling. Many of these processes have the potential to alter gene regulation across different trajectories and perhaps also orthogonally to the effects of cell type or differentiation stage.

To obtain a richer description of cellular identity and potentially resolve contrasts in cellular context beyond the discrete definition of cell type and differentiation stage, we performed topic modeling of HDC expression data. In this framework, transcriptional variation is decomposed into a fixed number of "cellular topics" represented by functional gene modules. We used FastTopics to identify 10 topics in the HDC dataset, aggregating cells into pseudocells (small clusters of transcriptionally similar cells) to mitigate noise and improve computational efficiency (STAR Methods).[41] Topic discovery was consistent across multiple resolutions of pseudocell aggregation (STAR Methods; Figure S5).

Several topics appeared to overlap with cell type and/or germ layer labels (Figure 4A), essentially recapitulating the results of cell type annotation: topic 1 was heavily loaded across endoderm-derived cell types, topic 2 across mesoderm-derived cell types, topic 5 on glial cells, and topic 6 on ectoderm-derived cell types. Two more topics encode developmental stage-specific information: topic 4 is most highly loaded on pluripotent cells and topic 7 is most heavily loaded on cells at intermediate stages of neuronal differentiation (Figure S6). This reinforces that cell type identity and developmental stage are the primary drivers of transcriptional variation, as mentioned above. Other topics, however, appeared to stratify cells based on gene programs that are less dependent on cell type or trajectory. For example, gene set enrichment analysis suggests that topic 8 appears to track the signature of the cell cycle in our data (Table S11). We confirmed this by directly estimating the cell cycle phase for all pseudocells (Figure 4B).[42] Topic 10 corresponds to a ciliary gene program that is shared across many cell types represented in this dataset.

We next sought to identify genetic variants with regulatory effects that are specific to the expanded set of cellular processes that were captured as topics (i.e., topic eQTLs). We limited this analysis to the 8 topics described above, as these were the most interpretable (STAR Methods).[43] We used CellRegMap to map topic eQTLs.[44] Instead of assessing each topic in isolation, CellRegMap jointly considers all linear combinations of topics, enabling us to simultaneously test for a wide range of genetic interactions with the cellular environment. We note that while varying the number of topics in the model involves trade-offs between expressivity and interpretability, the topic eQTLs detected in this analysis were consistently replicated when varying this hyperparameter (STAR Methods; Figure S7).
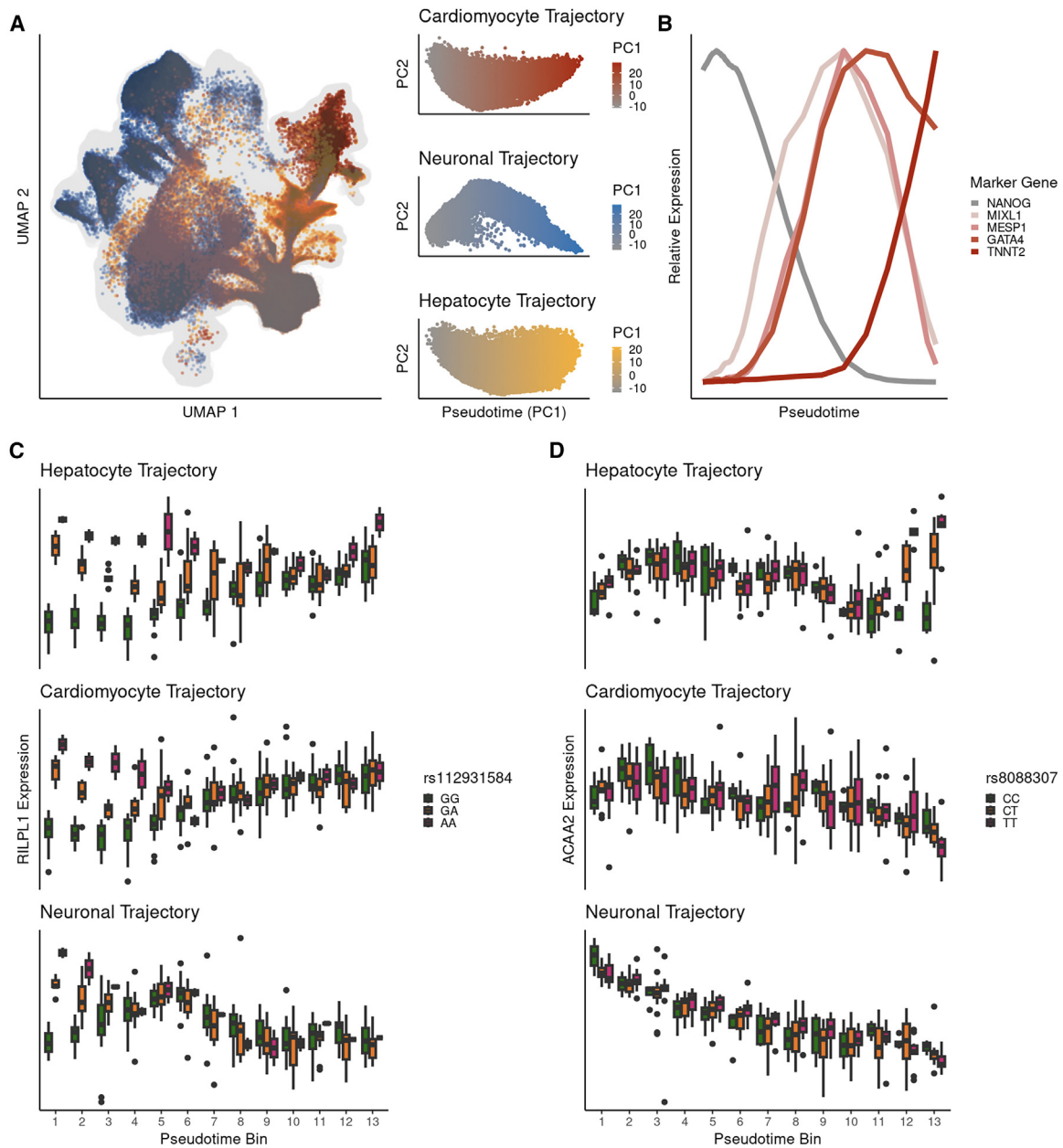
Figure 3. Dynamic eQTL calling along all three germ layers

(A) Trajectory isolation extracts cells mapped to three differentiation trajectories from the full dataset (left). The first expression principal component in each trajectory is used as a measure of pseudotime (right).

(B) Relative (min-max normalized) expression of 5 key marker genes displaying sequential expression with respect to pseudotime.

(C) Early dynamic eQTLs for the gene *RILPL1* shared across all three differentiation trajectories (center line, median normalized expression; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers).

(D) Late dynamic eQTLs for the gene *ACAA2* specific to the hepatocyte differentiation trajectory (top).

Since this powerful testing scheme is computationally expensive, we limited the topic interaction testing to the 77,550 eQTLs identified in the mash analysis. We identified a total of 157 genes with a topic eQTL (Table S12). For example, *DNA2* has a topic eQTL whose effect is correlated with the apparent cell cycle topic (topic 8; Figures 4C–4E). *DNA2* encodes a helicase protein involved in maintaining mitochondrial and nuclear DNA stability during DNA replication and repair. Visualizing this eQTL with the more clearly defined cell cycle phase inferred by tricycle[42] shows that this regulatory effect is largest during S phase, when DNA replication occurs. Another compelling example is the topic eQTL associated with *AXDND1*, with an effect that is correlated with the ciliary topic (topic 10; Figure S8). *AXDND1* is thought to encode a component of axonemal dyneins, which
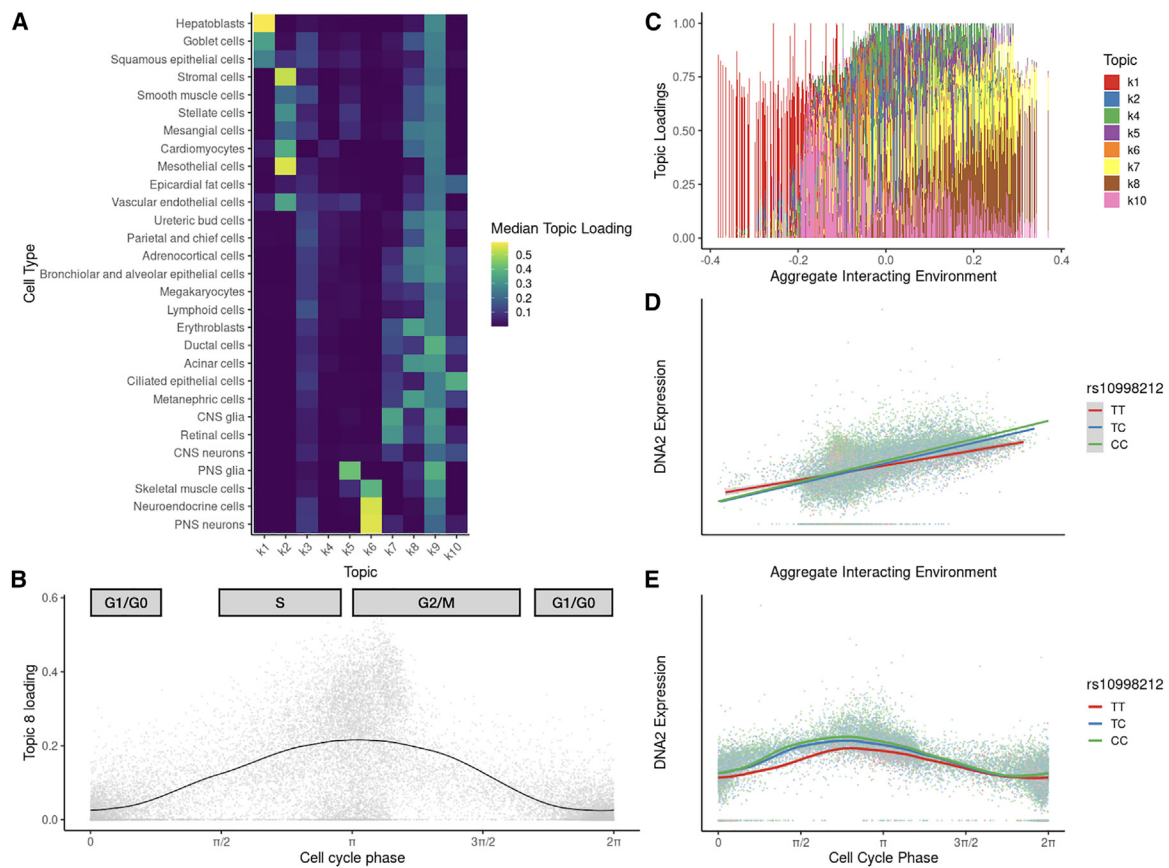
Figure 4. Topic eQTL calling

(A) Heatmap of median topic loadings in each cell type: while some topics are specific to one or a few cell types, others are associated with processes broadly shared across cell types.

(B) Topic 8 is associated with cell cycle phase, with the highest loadings on pseudocells in G2/M stage (phase). Solid curve is LOESS (locally estimated scatterplot smoothing) curve fit to all pseudocells.

(C) Structure plot of topic loadings. Each vertical bar shows topic loadings for a single pseudocell, sorted on the x axis by the aggregate signal of the cellular environment (aggregate interacting environment) for the topic interaction effect at rs10998212.

(D) Scatter plot of *DNA2* expression level versus the aggregate interacting environment for the topic interaction effect at rs10998212. Solid lines are linear regression lines for each genotype group.

(E) Visualizing the effect with respect to directly estimated cell cycle phase offers a more interpretable view of the regulatory dynamics, with maximal effect occurring in S phase. Solid curves are periodic LOESS curves fit for each genotype group.

drive ciliary beating to enable cell motility and extracellular fluid flow.[40,45] This analysis provides granular functional insight that is obscured by the standard catalog-based eQTL mapping approach.

HDC eQTLs help reveal the functional context of GWAS loci

Building on the successful application of our system to study gene regulation across a range of underexplored contexts, we next analyzed its utility in elucidating the molecular basis of complex traits. We began by focusing on the expression data, using scDRS to evaluate the relevance of the HDC cellular contexts to 40 traits with distinct biological underpinnings.[37] This analysis revealed significant associations between multiple trait-cell type pairs, including psychiatric traits such as schizophrenia and major depressive disorder linked to HDC neurons,

metabolic traits such as low-density lipoprotein (LDL) cholesterol levels associated with HDC hepatoblasts, and cardiovascular traits like coronary artery disease and diastolic blood pressure (DBP) related to HDC endothelial and stromal cells (Figure 5A).

To investigate whether genetic effects specific to HDC contexts could illuminate disease-associated loci that have not shown regulatory potential in adult tissue datasets, we shifted our focus from expression-based scDRS to HDC eQTLs. Using schizophrenia,[46] DBP,[47] and LDL cholesterol[48] as example traits, we confirmed that HDC eQTLs showed a higher disease risk than random variants, consistent with patterns observed in eQTLs from primary tissue samples (Figure S9). Notably, several of these eQTLs were previously uncharacterized: 7 of 18 genes (39%) with a schizophrenia-associated HDC eQTL, 25 of 93 (27%) with a DBP-associated HDC eQTL,
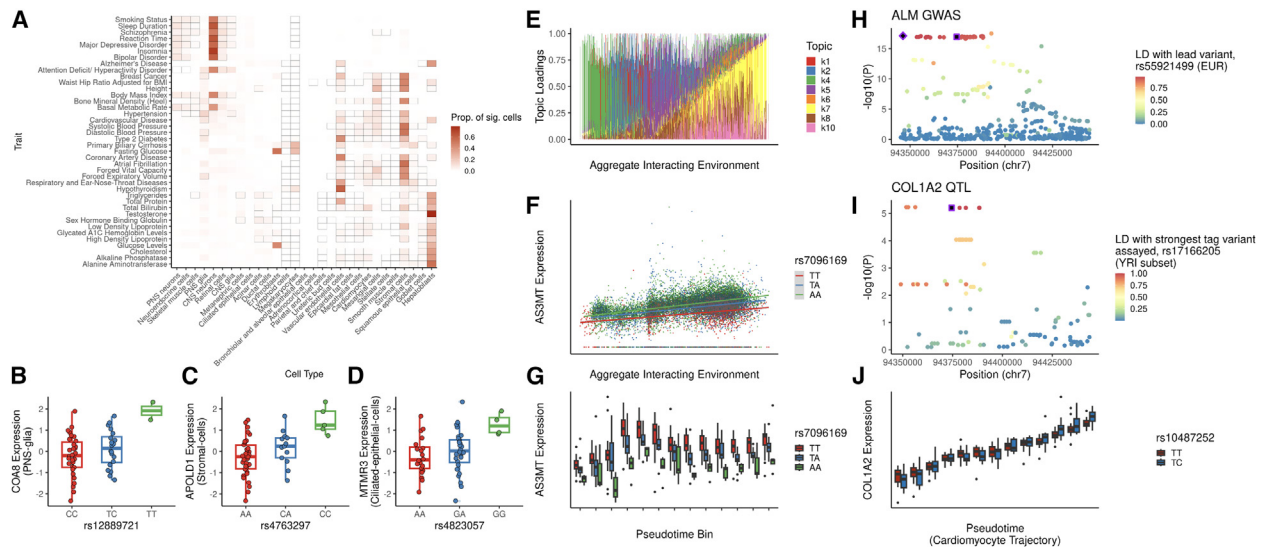
Figure 5. Exploring regulatory impacts of GWAS loci with HDCs

(A) Cell type-level disease relevance scores across 40 traits. Color depicts proportion of significantly associated cells for a trait belonging to a cell type, and box indicates significance at FDR 0.1 across all cell type-trait pairs.

(B–D) Trait-associated eQTLs not overlapping a significant variant-gene pair in GTEx: (B) a schizophrenia-associated eQTL for the gene *COA8* discovered in HDC PNS glial cells (center line, median normalized expression; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, individual observations); (C) a diastolic blood pressure-associated eQTL for *APOLD1* discovered in HDC stromal cells; and (D) an LDL-cholesterol-associated eQTL for *MTMR3* discovered in HDC ciliated epithelial cells.

(E) A topic-interaction eQTL at schizophrenia-associated locus rs7096169, a topic eQTL for *AS3MT*, draws the greatest contrast between the two endpoints of neuronal differentiation (topic 4, green, highest in pluripotent cells, and topic 6, purple, highest in neuronal cells), and the intermediate stages (topic 7, yellow).

(F) CellRegMap detects a significant interaction between genotype and this aggregate interacting environment on expression levels of the gene *AS3MT*.

(G) Visualizing this effect with respect to the more intuitive pseudotime measurement of neuronal differentiation state clarifies the nonlinear dynamic effect of cell state on the eQTL.

(H) Dynamic eQTL over the course of cardiomyocyte differentiation at an appendicular lean mass (ALM)-associated genetic variant.

(I) Locus plot displaying significance of association with ALM; points are colored by LD with the lead variant, shown with black diamond. As this variant has low frequency in the YRI panel used in the present study and was not tested for QTL effects, the assayed variant in strongest LD with the lead is additionally highlighted in the black square.

(J) Locus plot displaying significance of the interaction between genotype and pseudotime along the cardiomyocyte trajectory on *COL1A2* expression.

and 27 of 89 (30%) with an LDL-associated eQTL did not overlap with or tag (at linkage disequilibrium [LD] $R^2 \geq 0.5$) any significant eQTLs identified by GTEx. Examples include a schizophrenia-associated variant with an eQTL for *COA8*, a gene implicated in childhood neurodegenerative disease; a DBP-associated variant with an eQTL for *APOLD1*, which is involved in vascular function; and an LDL-associated variant with an eQTL for *MTMR3*, a gene involved in lipid metabolism (Figures 5B–5D).

Beyond mapping of individual genetic variants to candidate target genes, a more thorough characterization of regulatory dynamics offers insight into when and where the effects of a genetic variant may be most important. For example, rs7096169 is a known eQTL for *AS3MT* in several tissues, as well as a genome-wide significant schizophrenia risk locus. In HDCs, we found that this eQTL displayed a topic interaction effect, with the largest effect found in a topic associated with intermediate stages of neuronal development (topic 7) rather than either of the topics associated with an endpoint of neuronal development (topics 4 and 6; Figures 5E and 5F). We can clearly observe this variant's nonlinear dynamic regulatory effect when viewed with respect to pseudotime along the neuronal trajectory (Figure 5G).

While this triangulation of variant, gene, and context has previously been reported,[49,50] such efforts have historically required scanning large-scale databases of regulatory effects in adult or immune contexts and then establishing an *in vitro* platform to evaluate expression during differentiation into a single cell type selected *a priori*. The unification of an expansive set of cellular contexts at various stages of differentiation has the potential to accelerate this process, particularly when it is feasible to explore population-level genetic differences.

With this in mind, we next focused on novel interaction eQTLs (i.e., temporally dynamic eQTLs and topic eQTLs that are not found in GTEx, and did not tag a GTEx eQTL in any tissue at $R^2 \geq 0.5$). We used the Open Targets Genetics database to search for intersections between these novel regulatory effects and GWAS loci.[51] We found 62 genes with novel interaction eQTLs (ieGenes) displaying genome-wide significant association ($p \leq$ 5e−8) with at least one disease-associated locus: 26 neuronal dynamic ieGenes, 8 hepatocyte dynamic ieGenes, 10 cardiomyocyte dynamic ieGenes, and 24 topic ieGenes (Table S13). We highlight an example of a previously unknown dynamic eQTL for the collagen gene *COL1A2*, where the effect switches direction over the course

of cardiomyocyte differentiation, suggesting a regulatory element with contrasting effects over time, or the convergence of multiple dynamic effects in LD with the variant shown (Figures 5H–5J). This eQTL tags a variant that is associated with appendicular lean mass, a measure of the musculature in arms and legs.

DISCUSSION

The exploration of unguided HDCs in this study reveals an array of context-specific eQTLs that remained undetected in datasets such as GTEx, despite its comprehensive analysis of dozens of postmortem tissue types from hundreds of adult individuals. Our observations underscore the nuanced complexity of genetic regulation, which operates within highly specific cell types, states, and temporal contexts. By exploring cell types, trajectories, and programs that are difficult to sample *in vivo* and have never been studied from a population-level sample in humans, our work provides further support for the critical importance of context in understanding the regulatory mechanisms influencing disease.

We have demonstrated here that the HDC system offers valuable insights into gene regulation despite a lack of spatial organization mirroring *in vivo* tissues. Most of the cell type eQTLs discovered here overlap loci with regulatory function previously reported in primary tissue samples from the GTEx project. The HDC eQTLs that do not display this overlap demonstrate the expected clustering around the transcription start sites of genes but are enriched at known "blind spots" in existing resources: namely, the genes involved in a wide variety of developmental processes. Prioritizing flexibility enables us to activate a broader range of *cis*-regulatory elements that may be dormant in more accessible contexts.

Among the noteworthy findings from our HDC exploration is the identification of regulatory roles for dozens of disease-associated mutations that had not been linked to gene regulation in existing human eQTL studies. This highlights an important advantage of utilizing HDCs, which is the capacity to investigate a wide range of cellular contexts, some of which are rare or otherwise inaccessible in typical human samples. Enhanced by the detailed resolution of single-cell data and refined through trajectory inference and topic modeling, our approach enables the examination of not only distinct cell types and states but also subtler functional contexts that drive changes in gene regulation and the corresponding context-specific eQTLs.

The application of topic modeling to single-cell RNA-seq from HDCs has allowed us to traverse beyond traditional analyses confined by cell type or overall gene expression correlations. Topic modeling has revealed hidden layers of regulatory variation driven by dynamic processes such as cell division and ciliary activity that occur across multiple cell types and trajectories. By uncovering these additional dimensions of cellular identity, topic modeling has proven essential in identifying specific functional contexts that remained cryptic within the more conventional frameworks of single-cell classification.

To bridge the gaps in our understanding of the genetics of gene regulation, it is imperative to move beyond the static snapshots provided by adult tissues that have dominated eQTL studies in humans. HDCs facilitate this expansion by offering insights into the dynamic regulatory landscape of cellular differentiation. Looking forward, this system additionally presents the opportunity to further explore these diverse contexts under a wide range of chemical and genomic perturbations in order to better understand the role of gene-environment and gene-gene interactions. This expanded view of gene regulation offers a foundation to more deeply understand the genetics of complex traits, with the ultimate goal of accelerating the discovery of fundamental disease mechanisms and identifying potential therapeutic targets.

Limitations of the study
This study has certain limitations. First, while we have demonstrated the relevance of HDC cell types to various complex traits and shown concordance between HDC eQTLs and those identified in primary adult tissue samples, we recognize that the HDC system does not fully replicate *in vivo* development. Some contexts and context-specific regulatory effects identified in this study may not be present *in vivo*, and we do not expect to capture all developmental eQTLs within this model. Second, our current sample sizes constrain our ability to discover all eQTLs, fine-map causal variants underlying observed associations, and systematically analyze the contributions of specific cell types and states to disease heritability. Lastly, eQTL analyses tend to prioritize certain classes of functional variants,[11] which may differ systematically from trait-associated variants. Although these systematic differences are less likely to affect interaction eQTLs, association testing remains just one of several approaches for characterizing variant effects. As an *in vitro* system, HDCs can be employed with alternative assays, such as single-cell massively parallel reporter assays and CRISPR-based perturbation screens, which could offer complementary insights into the gene regulatory landscape across the diverse contexts explored in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, A.B. and Y.G.; methodology, J.M.P., K.R., K.B., A.B., and Y.G.; formal analysis, J.M.P., R.J., M.L., K.T., and K.B.; investigation, K.R. and K.B.; funding acquisition, A.B. and Y.G.; supervision, A.B. and Y.G.; writing – original draft, J.M.P., A.B., and Y.G.; writing– review & editing, J.M.P., K.R., R.J., M.L., K.B., K.T., A.B., and Y.G.

DECLARATION OF INTERESTS

K.R., A.B., and Y.G. are co-founders and equity holders of CellCipher. J.M.P. holds equity in CellCipher. K.R., K.B., and Y.G. are co-inventors on patent application 18067192 related to this work. A.B. is a stockholder in Alphabet, Inc., and has consulted for Third Rock Ventures.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Unguided HDC differentiation
  - Single-cell sequencing
  - Alignment, demultiplexing, and preliminary quality control
  - Cell type annotation
  - Dimensionality reduction, clustering, and visualization
  - Cell type eQTL calling
  - Multivariate adaptive shrinkage (MASH) eQTL calling
  - Trajectory isolation and pseudotime inference
  - Dynamic eQTL calling
  - Topic modeling
  - Topic eQTL calling
  - Single-cell disease relevance scoring
  - Trait-associated cell type eQTLs
  - Phenome-wide effects of interaction eQTLs
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

REFERENCES

1. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

2. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

3. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell *175*, 1701–1715.e16.

4. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

5. Kerimov, N., Hayhurst, J.D., Peikova, K., Manning, J.R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M.P., Kuzmin, I., Trevanion, S.J., et al. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. *53*, 1290–1299.

6. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310.

7. van der Wijst, M., de Vries, D.H., Groot, H.E., Trynka, G., Hon, C.C., Bonder, M.J., Stegle, O., Nawijn, M.C., Idaghdour, Y., van der Harst, P., et al. (2020). The single-cell eQTLGen consortium. Elife *9*, e52155. https://doi.org/10.7554/eLife.52155.

8. Yao, D.W., O'Connor, L.J., Price, A.L., and Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat. Genet. *52*, 626–633.

9. Umans, B.D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? Trends Genet. *37*, 109–124.

10. Connally, N.J., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., Cotsapas, C., Cassa, C.A., and Sunyaev, S.R. (2022). The missing link between genetic association and regulatory function. Elife *11*, e74970. https://doi.org/10.7554/eLife.74970.

11. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat. Genet. *55*, 1866–1875.

12. Findley, A.S., Monziani, A., Richards, A.L., Rhodes, K., Ward, M.C., Kalita, C.A., Alazizi, A., Pazokitoroudi, A., Sankararaman, S., Wen, X., et al. (2021). Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. Elife *10*, e67077. https://doi.org/10.7554/eLife.67077.

13. Arvanitis, M., Tayeb, K., Strober, B.J., and Battle, A. (2022). Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. Am. J. Hum. Genet. *109*, 223–239.

14. Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.-J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

15. Cuomo, A.S.E., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., et al. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nat. Commun. *11*, 810.

16. Jerber, J., Seaton, D.D., Cuomo, A.S.E., Kumasaka, N., Haldane, J., Steer, J., Patel, M., Pearce, D., Andersson, M., Bonder, M.J., et al. (2021). Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nat. Genet. *53*, 304–312.

17. Elorbany, R., Popp, J.M., Rhodes, K., Strober, B.J., Barr, K., Qi, G., Gilad, Y., and Battle, A. (2022). Single-cell sequencing reveals lineage-specific dynamic genetic regulation of gene expression during human cardiomyocyte differentiation. PLoS Genet. *18*, e1009666.

18. Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J.I., Baglaenko, Y., Suliman, S., Price, A.L., et al. (2022). Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature *606*, 120–128.

19. Cuomo, A.S.E., Nathan, A., Raychaudhuri, S., MacArthur, D.G., and Powell, J.E. (2023). Single-cell genomics meets human genetics. Nat. Rev. Genet. *24*, 535–549.

20. Kang, J.B., Shen, A.Z., Sakaue, S., Luo, Y., Gurajala, S., Nathan, A., Rumker, L., Aguiar, V.R.C., Valencia, C., Lagattuta, K., et al. (2023).

Mapping the dynamic genetic regulatory architecture of HLA genes at single-cell resolution. Preprint at medRxiv, 2023.03.14.23287257. https://doi.org/10.1101/2023.03.14.23287257.

21. Strober, B.J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., and Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. Science 364, 1287–1290.

22. Itskovitz-Eldor, J., Schuldiner, M., Karsenti, D., Eden, A., Yanuka, O., Amit, M., Soreq, H., and Benvenisty, N. (2000). Differentiation of human embryonic stem cells into embryoid bodies comprising the three embryonic germ layers. Mol. Med. 6, 88–95.

23. Brickman, J.M., and Serup, P. (2017). Properties of embryoid bodies. Wiley Interdiscip. Rev. Dev. Biol. 6, e259. https://doi.org/10.1002/wdev.259.

24. Rhodes, K., Barr, K.A., Popp, J.M., Strober, B.J., Battle, A., and Gilad, Y. (2022). Human embryoid bodies as a novel system for genomic studies of functionally diverse cell types. Elife 11, e71361. https://doi.org/10.7554/eLife.71361.

25. Barr, K.A., Rhodes, K.L., and Gilad, Y. (2023). The relationship between regulatory changes in cis and trans and the evolution of gene expression in humans and chimpanzees. Genome Biol. 24, 207.

26. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science 370, eaba7721. https://doi.org/10.1126/science.aba7721.

27. Müller, F.-J., Laurent, L.C., Kostka, D., Ulitsky, I., Williams, R., Lu, C., Park, I.-H., Rao, M.S., Shamir, R., Schwartz, P.H., et al. (2008). Regulatory networks define phenotypic classes of human stem cell lines. Nature 455, 401–405.

28. Banovich, N.E., Li, Y.I., Raj, A., Ward, M.C., Greenside, P., Calderon, D., Tung, P.Y., Burnett, J.E., Myrthil, M., Thomas, S.M., et al. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. 28, 122–131.

29. Bonder, M.J., Smail, C., Gloudemans, M.J., Frésard, L., Jakubosky, D., D'Antonio, M., Li, X., Ferraro, N.M., Carcamo-Orive, I., Mirauta, B., et al. (2021). Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. Nat. Genet. 53, 313–321.

30. Cuomo, A.S.E., Alvari, G., Azodi, C.B., single-cell eQTLGen consortium; McCarthy, D.J., and Bonder, M.J. (2021). Optimizing expression quantitative trait locus mapping workflows for single-cell studies. Genome Biol. 22, 188.

31. Murphy, A.E., and Skene, N.G. (2022). A balanced measure shows superior performance of pseudobulk methods in single-cell RNA-sequencing analysis. Nat. Commun. 13, 7851.

32. Zhou, H.J., Li, L., Li, Y., Li, W., and Li, J.J. (2022). PCA outperforms popular hidden variable inference methods for molecular QTL mapping. Genome Biol. 23, 210.

33. Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Stat. 31, 2013–2035.

34. Nguyen, J.P., Arthur, T.D., Fujita, K., Salgado, B.M., Donovan, M.K.R., iPSCORE Consortium; Matsui, H., Kim, J.H., D'Antonio-Chronowska, A., D'Antonio, M., and Frazer, K.A. (2023). eQTL mapping in fetal-like pancreatic progenitor cells reveals early developmental insights into diabetes risk. Nat. Commun. 14, 6928.

35. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. 51, 187–195.

36. Wilson, G.R., Sunley, J., Smith, K.R., Pope, K., Bromhead, C.J., Fitzpatrick, E., Di Rocco, M., van Steensel, M., Coman, D.J., Leventer, R.J., et al. (2014). Mutations in SH3PXD2B cause Borrone dermato-cardioskeletal syndrome. Eur. J. Hum. Genet. 22, 741–747.

37. Zhang, M.J., Hou, K., Dey, K.K., Sakaue, S., Jagadeesh, K.A., Weinand, K., Taychameekiatchai, A., Rao, P., Pisco, A.O., Zou, J., et al. (2022). Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. Nat. Genet. 54, 1572–1580.

38. Novembre, J., and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40, 646–649.

39. Burridge, P.W., Matsa, E., Shukla, P., Lin, Z.C., Churko, J.M., Ebert, A.D., Lan, F., Diecke, S., Huber, B., Mordwinkin, N.M., et al. (2014). Chemically defined generation of human cardiomyocytes. Nat. Methods 11, 855–860.

40. UniProt Consortium (2023). UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531.

41. Carbonetto, P., Sarkar, A., Wang, Z., and Stephens, M. (2021). Nonnegative matrix factorization algorithms greatly improve topic model fits. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.13440.

42. Zheng, S.C., Stein-O'Brien, G., Augustin, J.J., Slosberg, J., Carosso, G.A., Winer, B., Shin, G., Bjornsson, H.T., Goff, L.A., and Hansen, K.D. (2022). Universal prediction of cell-cycle position using transfer learning. Genome Biol. 23, 41.

43. Carbonetto, P., Luo, K., Sarkar, A., Hung, A., Tayeb, K., Pott, S., and Stephens, M. (2023). GoM DE: interpreting structure in sequence count data with differential expression analysis allowing for grades of membership. Genome Biol. 24, 236.

44. Cuomo, A.S.E., Heinen, T., Vagiaki, D., Horta, D., Marioni, J.C., and Stegle, O. (2022). CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. Mol. Syst. Biol. 18, e10663.

45. Walton, T., Wu, H., and Brown, A. (2021). Structure of a microtubule-bound axonemal dynein. Nat. Commun. 12, 477.

46. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604, 502–508.

47. Keaton, J.M., Kamali, Z., Xie, T., Vaez, A., Williams, A., Goleva, S.B., Ani, A., Evangelou, E., Hellwege, J.N., Yengo, L., et al. (2024). Genome-wide analysis in over 1 million individuals of European ancestry yields improved polygenic risk scores for blood pressure traits. Nat. Genet. 56, 778–791.

48. Graham, S.E., Clarke, S.L., Wu, K.-H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature 600, 675–679.

49. Washer, S.J., Flynn, R., Oguro-Ando, A., Hannon, E., Burrage, J., Jeffries, A., Mill, J., and Dempster, E.L. (2022). Functional characterization of the schizophrenia associated gene AS3MT identifies a role in neuronal development. Am. J. Med. Genet. B Neuropsychiatr. Genet. 189, 151–162.

50. Li, M., Jaffe, A.E., Straub, R.E., Tao, R., Shin, J.H., Wang, Y., Chen, Q., Li, C., Jia, Y., Ohi, K., et al. (2016). A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. Nat. Med. 22, 649–656.

51. Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. (2021). Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res. 49, D1311–D1320.

52. Genome Reference Consortium (2013). Genome assembly GRCh38. NCBI. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/.

53. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8, 14049.

54. Huang, Y., McCarthy, D.J., and Stegle, O. (2019). Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biol. 20, 273.

55. McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179–1186.

56. Cortal, A., Martignetti, L., Six, E., and Rausell, A. (2021). Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. Nat. Biotechnol. *39*, 1095–1102.

57. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

58. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

59. Taylor-Weiner, A., Aguet, F., Haradhvala, N.J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E.M., and Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. Genome Biol. *20*, 228.

60. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

61. Willwerscheid, J. (2021). Empirical Bayes Matrix Factorization: Methods and Applications. PhD thesis (University of Chicago).

62. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058.

63. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

64. 10× Genomics (2024). Cell Ranger (10× Genomics). https://www.10xgenomics.com/support/software/cell-ranger.

65. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

66. International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796.

67. Booeshaghi, A.S., Hallgrímsdóttir, I.B., Gálvez-Merchán, Á., and Pachter, L. (2022). Depth normalization for single-cell genomics count data. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.06.490859.

68. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

69. Murtagh, F., and Legendre, P. (2011). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. Preprint at arXiv. https://doi.org/10.48550/arXiv.1111.6285.

70. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

71. Zou, Y. (2021). Bayesian Variable Selection from Summary Data, with Application to Joint Fine-Mapping of Multiple Traits. PhD thesis (University of Chicago).

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

73. Prummel, K.D., Nieuwenhuize, S., and Mosimann, C. (2020). The lateral plate mesoderm. Development *147*, dev175059.

74. Yamashita, J., Itoh, H., Hirashima, M., Ogawa, M., Nishikawa, S., Yurugi, T., Naito, M., Nakao, K., and Nishikawa, S. (2000). Flk1-positive cells derived from embryonic stem cells serve as vascular progenitors. Nature *408*, 92–96.

75. Yamashita, J.K., Takano, M., Hiraoka-Kanie, M., Shimazu, C., Peishi, Y., Yanagi, K., Nakano, A., Inoue, E., Kita, F., and Nishikawa, S.-I. (2005). Prospective identification of cardiac progenitors by a novel single cell-based cardiomyocyte induction. Faseb. J. *19*, 1534–1536.

76. Litviňuková, M., Talavera-López, C., Maatz, H., Reichart, D., Worth, C.L., Lindberg, E.L., Kanda, M., Polanski, K., Heinig, M., Lee, M., et al. (2020). Cells of the adult human heart. Nature *588*, 466–472.

77. Takada, S., Stark, K.L., Shea, M.J., Vassileva, G., McMahon, J.A., and McMahon, A.P. (1994). Wnt-3a regulates somite and tailbud formation in the mouse embryo. Genes Dev. *8*, 174–189.

78. Uosaki, H., and K, J. (2011). Chemicals Regulating Cardiomyocyte Differentiation. In Embryonic Stem Cells: The Hormonal Regulation of Pluripotency and Embryogenesis, C. Atwood, ed. (InTechOpen). https://doi.org/10.5772/589.

79. Balafkan, N., Mostafavi, S., Schubert, M., Siller, R., Liang, K.X., Sullivan, G., and Bindoff, L.A. (2020). A method for differentiating human induced pluripotent stem cells toward functional cardiomyocytes in 96-well microplates. Sci. Rep. *10*, 18498.

80. Carpenedo, R.L., Kwon, S.Y., Tanner, R.M., Yockell-Lelièvre, J., Choey, C., Doré, C., Ho, M., Stewart, D.J., Perkins, T.J., and Stanford, W.L. (2019). Transcriptomically guided mesendoderm induction of human pluripotent stem cells using a systematically defined culture scheme. Stem Cell Rep. *13*, 1111–1125.

81. The International Stem Cell Initiative (2007). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat. Biotechnol. *25*, 803–816.

82. Green, M.D., Chen, A., Nostro, M.-C., d'Souza, S.L., Schaniel, C., Lemischka, I.R., Gouon-Evans, V., Keller, G., and Snoeck, H.-W. (2011). Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. Nat. Biotechnol. *29*, 267–272.

83. Han, L., Chaturvedi, P., Kishimoto, K., Koike, H., Nasr, T., Iwasawa, K., Giesbrecht, K., Witcher, P.C., Eicher, A., Haines, L., et al. (2020). Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. Nat. Commun. *11*, 4158.

84. Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J.R., and Zaret, K.S. (1996). Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. Genes Dev. *10*, 1670–1682.

85. Costa, R.H., Grayson, D.R., and Darnell, J.E., Jr. (1989). Multiple hepatocyte-enriched nuclear factors function in the regulation of transthyretin and alpha 1-antitrypsin genes. Mol. Cell Biol. *9*, 1415–1425.

86. Lai, E., Prezioso, V.R., Smith, E., Litvin, O., Costa, R.H., and Darnell, J.E., Jr. (1990). HNF-3A, a hepatocyte-enriched transcription factor of novel structure is regulated transcriptionally. Genes Dev. *4*, 1427–1436.

87. Sladek, F.M., Zhong, W.M., Lai, E., and Darnell, J.E., Jr. (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. Genes Dev. *4*, 2353–2365.

88. Cai, J., Zhao, Y., Liu, Y., Ye, F., Song, Z., Qin, H., Meng, S., Chen, Y., Zhou, R., Song, X., et al. (2007). Directed differentiation of human embryonic stem cells into functional hepatic cells. Hepatology *45*, 1229–1239.

89. Lendahl, U., Zimmerman, L.B., and McKay, R.D. (1990). CNS stem cells express a new class of intermediate filament protein. Cell *60*, 585–595.

90. Zhang, X., Huang, C.T., Chen, J., Pankratz, M.T., Xi, J., Li, J., Yang, Y., Lavaute, T.M., Li, X.-J., Ayala, M., et al. (2010). Pax6 is a human neuroectoderm cell fate determinant. Cell Stem Cell *7*, 90–100.

91. Thomson, M., Liu, S.J., Zou, L.-N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875–889.

92. Li, X.-J., Du, Z.-W., Zarnowska, E.D., Pankratz, M., Hansen, L.O., Pearce, R.A., and Zhang, S.-C. (2005). Specification of motoneurons from human embryonic stem cells. Nat. Biotechnol. *23*, 215–221.

93. Shimojo, H., Ohtsuka, T., and Kageyama, R. (2011). Dynamic expression of notch signaling genes in neural stem/progenitor cells. Front. Neurosci. *5*, 78.

94. Seo, S., Lim, J.-W., Yellajoshyula, D., Chang, L.-W., and Kroll, K.L. (2007). Neurogenin and NeuroD direct transcriptional targets and their regulatory enhancers. EMBO J. *26*, 5093–5108.

95. Noisa, P., Lund, C., Kanduri, K., Lund, R., Lähdesmäki, H., Lahesmaa, R., Lundin, K., Chokechuwattanalert, H., Otonkoski, T., Tuuri, T., and Raivio, T. (2014). Notch signaling regulates the differentiation of neural crest from human pluripotent stem cells. J. Cell Sci. *127*, 2083–2094.

96. Liang, X., Song, M.-R., Xu, Z., Lanuza, G.M., Liu, Y., Zhuang, T., Chen, Y., Pfaff, S.L., Evans, S.M., and Sun, Y. (2011). Isl1 is required for multiple aspects of motor neuron development. Mol. Cell. Neurosci. *47*, 215–222.

97. Nadal-Nicolás, F.M., Jiménez-López, M., Sobrado-Calvo, P., Nieto-López, L., Cánovas-Martínez, I., Salinas-Navarro, M., Vidal-Sanz, M., and Agudo, M. (2009). Brn3a as a marker of retinal ganglion cells: qualitative and quantitative time course studies in naive and optic nerve-injured retinas. Invest. Ophthalmol. Vis. Sci. *50*, 3860–3868.

98. des Portes, V., Francis, F., Pinard, J.-M., Desguerre, I., Moutard, M.-L., Snoeck, I., Meiners, L.C., Capron, F., Cusmai, R., Ricci, S., et al. (1998). Doublecortin is the major gene causing X-linked subcortical laminar heterotopia (SCLH). Hum. Mol. Genet. *7*, 1063–1070.

99. Guo, Z., Zhang, L., Wu, Z., Chen, Y., Wang, F., and Chen, G. (2014). In vivo direct reprogramming of reactive glial cells into functional neurons after brain injury and in an Alzheimer's disease model. Cell Stem Cell *14*, 188–202.

100. Englund, C., Fink, A., Lau, C., Pham, D., Daza, R.A.M., Bulfone, A., Kowalczyk, T., and Hevner, R.F. (2005). Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex. J. Neurosci. *25*, 247–251.

101. Cecchi, C. (2002). Emx2: a gene responsible for cortical development, regionalization and area specification. Gene *291*, 1–9.

102. Weihe, E., Schäfer, M.K., Erickson, J.D., and Eiden, L.E. (1994). Localization of vesicular monoamine transporter isoforms (VMAT1 and VMAT2) to endocrine cells and neurons in rat. J. Mol. Neurosci. *5*, 149–164.

103. Britanova, O., de Juan Romero, C., Cheung, A., Kwan, K.Y., Schwark, M., Gyorgy, A., Vogel, T., Akopov, S., Mitkovski, M., Agoston, D., et al. (2008). Satb2 is a postmitotic determinant for upper-layer neuron specification in the neocortex. Neuron *57*, 378–392.

104. Leid, M., Ishmael, J.E., Avram, D., Shepherd, D., Fraulob, V., and Dollé, P. (2004). CTIP1 and CTIP2 are differentially expressed during mouse embryogenesis. Gene Expr. Patterns *4*, 733–739.

105. Walker, K.L., Yoo, H.K., Undamatla, J., and Szaro, B.G. (2001). Loss of neurofilaments alters axonal growth dynamics. J. Neurosci. *21*, 9655–9666.

106. Mullen, R.J., Buck, C.R., and Smith, A.M. (1992). NeuN, a neuronal specific nuclear protein in vertebrates. Development *116*, 201–211.

107. Ferreira, A., and Caceres, A. (1992). Expression of the class III beta-tubulin isotype in developing neurons in culture. J. Neurosci. Res. *32*, 516–529.

108. Matus, A., Bernhardt, R., and Hugh-Jones, T. (1981). High molecular weight microtubule-associated proteins are preferentially associated with dendritic microtubules in brain. Proc. Natl. Acad. Sci. USA *78*, 3010–3014.

109. Wold, H. (1975). Path models with latent variables: The NIPALS approach. In Quantitative Sociology, H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon, and V. Capecchi, eds. (Academic Press), pp. 307–357.

110. Davis, J.R., Fresard, L., Knowles, D.A., Pala, M., Bustamante, C.D., Battle, A., and Montgomery, S.B. (2016). An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. Am. J. Hum. Genet. *98*, 216–224.

111. Strober, B.J., Tayeb, K., Popp, J., Qi, G., Gordon, M.G., Perez, R., Ye, C.J., and Battle, A. (2024). SURGE: uncovering context-specific genetic-regulation of gene expression from single-cell RNA sequencing using latent-factor models. Genome Biol. *25*, 28.

112. The Gene Ontology Consortium; Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*, iyad031. https://doi.org/10.1093/genetics/iyad031.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
10X Genomics 3′ scRNA-seq v3.1	10x Genomics	PN-1000268
High Sensitivity DNA Kit	Agilent	Part Number: 5067-4626
Deposited data		
Human Reference genome GRCh38	Genome Reference Consortium[52]	GRCh38: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/
Single-Cell RNA-seq data from human iPSC lines NA19160, NA18511, and NA18858	Rhodes et al.[24]	GEO: GSE178274
Fetal Cell Atlas	Cao et al.[26]	Fetal Cell Atlas: https://atlas.brotmanbaty.org/bbi/human-gene-expression-during-development/
Single-cell RNA-seq from cardiomyocyte differentiation timecourse	Elorbany et al.[17]	GEO: GSE175634
Curated MAGMA gene sets	Zhang et al.[37]	FigShare: https://figshare.com/articles/dataset/scDRS_data_release_092121/16664080?file=30853708
GTEx v8	The Genotype-Tissue Expression Project	GTEx Portal: https://gtexportal.org/home/
Schizophrenia GWAS summary statistics	Trubetskoy et al.[46]	PGC: https://pgc.unc.edu/for-researchers/download-results/scz/
Diastolic blood pressure GWAS summary statistics	Keaton et al.[47]	GWAS Catalog: https://www.ebi.ac.uk/gwas/studies/GCST90310295
LDL cholesterol GWAS summary statistics	Graham et al.[48]	Global Lipid Genetics Consortium: https://csg.sph.umich.edu/willer/public/glgc-lipids2021/
Aggregated human GWAS data	Open Targets Genetics	Open Targets Genetics Database: https://genetics.opentargets.org/
Raw single-cell gene expression data	This Paper	SRA: SRP496700
Preprocessed single-cell gene expression data	This Paper	GEO: GSE266500
Experimental models: Cell lines		
53 human iPSC lines	Banovich et al.	Original study: https://pubmed.ncbi.nlm.nih.gov/29208628/
Software and algorithms		
10x Genomics Cell Ranger v6.1.2	Zheng et al.[53]	cellranger: https://www.10xgenomics.com/support/software/cell-ranger/latest
VireoSNP v0.5.6	Huang et al.[54]	vireo: https://vireosnp.readthedocs.io/en/latest/
Scater v1.26.0	McCarthy et al.[55]	scater: https://bioconductor.org/packages/release/bioc/html/scater.html
Cell-ID v1.6.0	Cortal et al.[56]	Cell-ID: https://github.com/RausellLab/CelliD
Scanpy v1.9.1	Wolf et al.[57]	scanpy: https://scanpy.readthedocs.io/en/stable/
edgeR v3.40.0	Robinson et al.[58]	edgeR: https://bioconductor.org/packages/release/bioc/html/edgeR.html
TensorQTL v1.0.7	Taylor-Weiner et al.[59]	tensorQTL: https://github.com/broadinstitute/tensorqtl
Bedtools v2.27.1	Quinlan and Hall[60]	bedtools: https://bedtools.readthedocs.io/en/latest/index.html

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mashr v1.0.7	Urbut et al.	mashr: https://github.com/stephenslab/mashr
flashier v0.2.36	Willwerscheid et al.[61]	flashier: https://github.com/willwerscheid/flashier
scDRS v1.0.2	Zhang et al.[37]	scDRS: https://github.com/martinjzhang/scDRS
scvi-tools v0.20.0	Lopez et al.[62]	scvi-tools: https://docs.scvi-tools.org/en/stable/user_guide/models/scvi.html
FastTopics v0.6-142	Carbonetto et al.[41,43]	FastTopics: https://github.com/stephenslab/fastTopics
Plink v1.90	Purcell et al.[63]	PLINK: https://zzz.bwh.harvard.edu/plink/
CellRegMap v0.0.3	Cuomo et al.[44]	CellRegMap: https://limix.github.io/CellRegMap/
Original Code	This Paper	Github: https://github.com/jmp448/hdcQTL/tree/v1.0.0 (and Zenodo: https://doi.org/10.5281/zenodo.14009254)
QTL Summary Statistics	This Paper	Zenodo: https://zenodo.org/records/13694387

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We use induced pluripotent stem cell lines of unrelated Yoruba individuals from Ibadan, Nigeria (YRI). 51 iPSC lines were differentiated and collected for this study. These were analyzed along with data from 3 cell lines from Rhodes et al. 2022. One line (NA18858) was collected in both studies, bringing the total number of cell lines analyzed in this study to 53. Metadata (including sex of cells) for each cell line is available in the cell metadata available on GEO (see data and code availability).

We maintained feeder-free iPSC cultures on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with StemFlex Medium (A3349401, Thermo Fisher Scientific) and Penicillin/Streptomycin (30,002 Cl, Corning). We grew cells in an incubator at 37°C, 5% CO2, and atmospheric O2. Every 3–5 days thereafter, we passaged cells to a new dish using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded cells with ROCK inhibitor Y-27632 (ab120129, Abcam).

METHOD DETAILS

Unguided HDC differentiation

HDCs were formed using a modified version of the STEMCELL Agrewell400 protocol which was previously described in Rhodes et al.[24] We coated wells of an Aggrewell 400 24-well plate (34414, STEMCELL) with anti-adherence rinsing solution (07010, STEMCELL). iPSCs were seeded into the Aggrewell 400 24-well plate at a density of 1,000 cells per microwell in Aggrewell EB Formation Medium (05893, STEMCELL) with ROCK inhibitor Y-27632 and Penicillin/Streptomycin. After 24 h, we replaced half of the spent media with fresh Aggrewell EB Formation Medium without ROCK inhibitor. 48 h after seeding the Aggrewell plate, we harvested EBs and moved them to an ultra-low attachment 6-well plate (CLS3471-24EA, Sigma) in E6 media (A1516401, ThermoFisher Scientific) with Penicillin/Streptomycin. We maintained HDCs in culture for an additional 19 days, replacing media with fresh E6 media every 48 h.

HDCs were dissociated for collection 21 days after formation. HDCs were dissociated by washing them with phosphate-buffered saline (Corning 21-040-CV), treating them with AccuMax (STEMCELL 7921) and incubating them at 37C for 15-40 min total. After the first 10 min in Accumax, we pipetted HDCs up and down with a wide-bore p1000 pipette tip for 30 s. Subsequently, we repeated pipetting with a standard p1000 pipette tip for 30 s every 5 min until HDCs were completely dissociated. We then quenched the dissociation by adding E6 media to cells and strained them through a 40 μm strainer (Fisherbrand 22-363-547). We resuspended cells in PBS with 0.04% bovine serum albumin and counted them. Early collection batches were counted using a TC20 Automated Cell Counter (450102, BioRad), and later batches were counted using a Countess II (AMQAF1000, invitrogen). We then mixed lines together in equal proportions prior to collection.

We differentiated two replicates of 51 iPSC lines across 17 batches consisting of 4–8 lines per batch. Each batch contained at least one line of each sex. In each batch, lines were formed and maintained in parallel. Each replicate represents a separate instance of HDC differentiation and will capture technical effects introduced during formation, maintenance, dissociation, and collection. We refer to these batches as "collection batches".

Single-cell sequencing

We generated scRNA-seq libraries using the 10X Genomics 3′ scRNA-seq v3.1 kit. Using the evenly pooled mix of lines from each collection batch, we loaded a 10x chip targeting 10,000 cells per lane of the 10x chip and loaded the same pool of cells across multiple lanes to recover 10,000 per individual. After cDNA amplification and cleanup, samples from each collection batch were stored at −20°C. Library preparation proceeded in larger batches composed of two or more collection batches (see Table S1). For example, in library preparation batch 1, all samples from collection batches 1 and 3 were performed in parallel. cDNA libraries for biological replicates were always processed in different library preparation batches. Libraries were sequenced on the NovaSeq in the University of Chicago Functional Genomics Core. We pooled samples for sequencing in 7 batches. These batches are distinct from library preparation batch and are composed of two or more samples from the same collection batch. Some samples from a single collection day were split across different "sequencing batches" (Table S1). We targeted a final sequencing depth of 100,000 reads per cell.

Alignment, demultiplexing, and preliminary quality control

We used Cellranger[53] to align samples to the human genome (GRCh38[52]) and to aggregate samples from all collections from this study as well as additional libraries collected previously that included 2 additional Yoruba individuals (NA19160, NA18511) and one individual represented in both collections (NA18858).[24,25,64] We used Vireo to demultiplex samples and assign droplets to individuals. We used previously collected and imputed genotypes for the included Yoruba individuals from the HapMap and 1000 Genomes Project.[54,65,66] We then filtered cells to remove droplets that Vireo identified as doublets and droplets that could not be confidently assigned to an individual. For the samples collected prior to this study, we kept only droplets representing Yoruba individuals (these samples originally included chimpanzee cells and cells from non-YRI humans). We further filtered cells to keep only those with less than 15% mitochondrial reads and with at least 2500 genes expressed. We also removed cells with very high total counts, keeping only those with less than 150,000 total counts. Finally, we filtered to genes expressed in at least 10 cells. This left a total of 909,536 cells and 35,324 genes for downstream analysis.

Cell type annotation

The fetal cell atlas contains a total of 77 cell type labels. To obtain marker gene sets for these 77 cell types, we subsampled over-abundant cell types to a maximum of 5,000 cells per cell type (regardless of the tissue of origin), and filtered genes to protein-coding genes. To assess classifier performance, we treated cell type labels assigned by the original fetal cell atlas paper as ground truth. We split the uniformly sampled data into training and testing subsets with equal numbers of cells. We preprocessed the training data using *scater*[55]: we first normalized cells by size factors, then log transformed the data, and selected highly variable features at an FDR threshold of 0.1. We then performed multiple correspondence analysis (MCA) to extract signature gene sets for each cell type using *Cell-ID*.[56] Annotation of the HDC data with these gene sets suggested that 64 of 77 cell types were represented with at least 5 cells present from at least 25 donors. However, this classifier displayed poor accuracy on held-out test data (Figure S1). Examination of the reference gene sets suggested that high similarity among related cell types (e.g., neuronal subtypes) likely compromised the classifier's performance.

In order to obtain a more limited set of maximally interpretable cell type labels, we first removed 12 cell types which were poorly characterized in the reference dataset, or attributed to potential contamination: AFP_ALB positive cells, CCL19_CCL21 positive cells, CLC_IL5RA positive cells, CSH1_CSH2 positive cells, ELF3_AGBL2 positive cells, MUC13_DMBT1 positive cells, PDE11A_FAM19A2 positive cells, PDE1C_ACSM3 positive cells, SATB2_LRRC7 positive cells, SKOR2_NPSR1 positive cells, SLC24A4_PEX5L positive cells, and SLC26A4_PAEP positive cells. We additionally removed placental cells from the reference which are unlikely to arise in the *in vitro* HDC system, as well as rare cell types represented by fewer than 500 cells for which it may be difficult to define a meaningful expression signature. We combined limbic system neurons, inhibitory interneurons, inhibitory neurons, excitatory neurons, unipolar brush cells, granule neurons, and Purkinje neurons under the label of central nervous system (CNS) neurons; microglia, astrocytes, and oligodendrocytes under CNS glia; visceral neurons and enteric nervous system (ENS) neurons under peripheral nervous system (PNS) neurons; ENS glia, satellite cells, and Schwann cells under PNS glia; retinal progenitors and Muller glia, amacrine cells, bipolar cells, ganglion cells, retinal pigment cells, horizontal cells, and photoreceptor cells under retinal cells; and chromaffin cells, Islet endocrine cells, neuroendocrine cells, and sympathoblasts under neuroendocrine cells. Immune cell types were also quite granularly defined in this reference, posing a challenge to a global cell type classifier. We therefore removed myeloid cells and hematopoietic stem cells from the reference, which are common ancestors of more specific cell types present in the reference, and merged thymocytes with lymphoid cells due to their shared lymphoid progenitor. Altogether, this left us with a refined set of 33 cell types. After once again subsampling overrepresented cell type labels to at most 5000 cells, and re-defining our set of highly variable genes using the same criteria as above, classification accuracy increased to nearly 90% (Figure S1).

Our final set of cell type signatures was obtained by applying the sampling, normalization, and embedding procedures described above to the merged training and test datasets. The gene set signatures used for this final classifier are available in Table S2.

To classify HDC cells using these gene sets, we first normalized our HDC cells as described in the previous section, subsetting to the same set of highly variable genes used for the fetal cell atlas embedding. We applied MCA, and used *Cell-ID* to generate cell type labels.

Dimensionality reduction, clustering, and visualization

To generate the UMAP embedding, we first identified 5000 highly variable genes using proportional fitting[67] and scanpy's default method for extracting highly variable genes proposed in.[68] We subset to these 5,000 genes to generate a 50-dimensional embedding of the data using a variational autoencoder applied to raw UMI counts with *scVI*.[62] To generate a 2-dimensional embedding of the data we computed a neighborhood graph based on this encoding and applied UMAP using *scanpy*'s default settings.[57] This UMAP embedding was used purely for the convenience of visualizing the full dataset in two dimensions, and did not influence cell type annotation, trajectory inference, topic modeling, or any QTL calling.

To group cel -types according to expression similarity, we aggregated all cells with the same cell type label into a pseudobulk sample by taking the sum of all UMI counts. We then filtered to the intersection of protein-coding genes and the fetal cell atlas highly variable gene set (see *Classifier Development* section), and applied TMM and log CPM normalization using the *edgeR* package.[58] We applied hierarchical clustering to these normalized pseudobulk samples for each cell type using Ward's method.[69]

Cell type eQTL calling

To perform eQTL calling, we aggregated cells from the same donor and cell type by taking the sum of all UMI counts per gene. We removed pseudobulk samples with fewer than 5 cells. Within each cell type, we filtered to genes with nonzero variance, and with a median expression level of at least 10 UMI counts per sample. We applied log CPM normalization to each sample, using TMM normalization factors computed separately in each cell type with the *edgeR* package.[58,70] We then applied an inverse normal transformation to each gene. We computed expression principal components to include as latent covariates for eQTL calling, as well as for quality control: principal component biplots were manually inspected to identify and remove outlier samples which may disrupt eQTL calling. Results of these filtering criteria are available in Table S3. Donor sex was additionally included as a covariate.

We tested all variants within 50kb of the corresponding gene's TSS, that had minor allele frequency of at least 0.1 among specifically the samples included in each cell-type-specific analysis (note that this set of donors varies between cell types as demonstrated in Figure 1D). Genotypes were centered and scaled separately in each cell type to maximize comparability of effect size estimates across cell types. (This does not influence significance tests for the cell type-by-cell type analysis but becomes relevant for the mash analysis described below). QTL calling was conducted with TensorQTL, using beta-approximated *p* values based on permutations to control for multiple testing burden at each gene.[59]

To generate a list of all significant variant-gene pairs, we followed the procedure described by the GTEx Consortium[4]: a genome-wide *p* value threshold was defined as the beta-approximated *p* value of the gene closest to the global q value cutoff of 0.05. This genome-wide threshold was used to define a nominal threshold for each gene based on the per-gene beta distribution estimated with TensorQTL, and all variants with a nominal *p* value below this gene-level threshold were considered significant (Table S5).

We used bedtools[60] and awk to identify variant-gene pairs intersecting the full set of significant eQTL variant-gene pairs in each GTEx tissue, using data from GTEx v8. We first intersected variants between the two sets of eQTLs, then further filtered to gene-variant pairs with matching Ensembl gene IDs.

Multivariate adaptive shrinkage (MASH) eQTL calling

To perform the mash analysis, we first estimated residual correlation structure (V) with mash's expectation maximization procedure, subsetting to a random subset of 25,000 gene-variant pairs that were tested in all cell types for efficiency.[71] Next, to estimate data-driven covariance matrices (U), we subset first to gene-variant pairs which were tested in all cell types, then to the strongest (largest absolute *Z* score) effect per gene, and finally to the top 2,000 strongest effects across all genes. We decomposed this set of 2,000 eQTL effects across cell types into non-negative latent factors through Empirical Bayes Matrix Factorization using flashier[61] as implemented in the mash package. We removed singleton components which only assigned weight to a single cell type. We included the four rank-1 covariance matrices generated from these latent factors, as well as their normalized sum, to fit the mash model. We additionally included a singleton component for each cell type as well as a component corresponding to shared effects across all cell types. We fit the mash model on a random subset of 50,000 gene-variant pairs that were tested in at least 10 contexts. After model fitting, we performed inference across all variant-gene pairs tested in at least 5 contexts.

Trajectory isolation and pseudotime inference

For each trajectory, we curated marker gene lists from multiple published directed differentiation protocols defining marker genes for each stage of differentiation.[17,72–108] We used scDRS to test for enrichment of the marker gene set compared to 50 background gene sets matched for mean and variance of library-size normalized, log-transformed single-cell expression,[37] filtering to cells with FDR $\leq$ 0.1 for inclusion in the isolated trajectory. Marker gene sets are available in Tables S7, S8, and S9.

We applied principal component analysis to library-size, log normalized single-cell gene expression data from each trajectory in isolation, and used this first principal component as a measure of pseudotime. To visualize trends in marker gene expression with respect to pseudotime, we adopted a sliding window approach used by Cuomo et al. (taking the average expression of 10% of cells in each window, sliding along pseudotime by 2.5% of cells),[15] followed by min-max normalization for each gene.

Dynamic eQTL calling

To mitigate the noise of single-cell data and leverage efficient and established tools for interaction eQTL calling, we aggregated cells into pseudobulk samples. We grouped cells into 15 pseudotime bins of equal range. In each trajectory, cells from bins 1 and 2 were combined as were cells from bins 14 and 15 to account for fewer cells and donors being represented at the extremes of the pseudotime distribution. Each pseudobulk sample contains the sum of UMI counts from all cells from a single donor in a single pseudotime bin.

In each trajectory, we dropped pseudobulk samples that consisted of less than 5 cells, omitted any donors represented in 10 or fewer pseudotime bins, and filtered to genes with non-zero pseudobulk expression in at least 10 samples and non-zero variance in expression across samples. This resulted in a total of 428 pseudobulk samples from a total of 35 cell lines for the cardiomyocyte trajectory, 383 samples from 31 cell lines in the hepatocyte trajectory, and 549 samples from 44 cell lines in the neuronal trajectory. We normalized pseudobulk expression and filtered tests as described for the cell type analysis.

As in previous work,[17,21] we used cell line PCA to infer latent covariates introducing broad differences in expression between cell lines over the course of a differentiation trajectory. We used the NIPALS algorithm[109] to perform cell line PCA with missing values, as not all donors contain 5 cells in each pseudotime bin. We included sex and the first 10 cell line PCs as covariates, as well as their interaction with pseudotime.

We use TensorQTL[59] to conduct dynamic eQTL calling, using the following model:

$$E_{ct} \sim N\left(\mu + \beta_1 G_c + \beta_2 \tau + \beta_3 PC_c^1 + ... + \beta_{12} PC_c^{10} + \beta_{13} S_c + \beta_{14} PC_c^1 \tau + ... + \beta_{23} PC_c^{10} \tau + \beta_{24} S_c \tau + \beta_{25} G_c \tau, \sigma\right)$$

Where c indexes donors and t indexes pseudotime bins. E_{ct} represents the normalized expression of the sample, G represents genotype, PC^k represents the kth cell line PC, S_c represents the sex of donor c, and τ represents median pseudotime value of all cells in the sample.

We perform inference on the interaction effect between genotype and pseudotime (β_{25}). We used eigenMT to adjust for multiple QTL tests conducted per gene.[110] We then apply the Benjamini-Hochberg procedure to the smallest adjusted p value per gene to control the genome-wide false discovery rate at a level of 0.1.

We generated a list of all significant dynamic eQTLs analogously to the procedure described for cell type eQTLs: a genome-wide adjusted p value threshold was defined as the EigenMT-adjusted p value of the gene closest to the global FDR threshold, and all variants with an adjusted p value below this threshold were included in the final set of significant dynamic eQTLs (Table S10).

To test for replication of these eQTLs in a separate dataset, we reprocessed the single-cell expression data collected in Elorbany et al. as described above, and used the *qvalue* package to estimate the π_1 replication rate for the strongest cardiomyocyte dynamic eQTL per gene detected in our study.[33]

We classified significant dynamic eQTLs as early, late, or switch categories as in previous work.[17,21] An eQTL is classified as early if the effect size decreased over time, late if it increased over time, and switch if the sign of the effect changed over time. We used the fitted linear model generated by TensorQTL to estimate predicted effect sizes at the endpoints of the trajectory's pseudotime range, and compare these predictions to classify each eQTL. If the effect size remains the same, the variant is classified as early if the magnitude decreases or late if the magnitude increases. If the effect size flips and the predicted effect at both ends has a magnitude exceeding a threshold of 1, it is called a switch effect.

Topic modeling

To improve computational efficiency and once again mitigate the noise of single-cell data, we aggregated cells into pseudocells before applying topic modeling in an approach similar to that used by Strober et al.[111] To avoid over-representation from donors with outlier cell counts, for donors with cell counts over 1.5 standard deviations above the median we randomly subsampled without replacement to the median number (16,839) of cells. We also removed cells collected in a previous study[24] due to differences in cell depth that disrupted topic modeling, leaving 51 of 53 donors (see 'Unguided HDC Differentiation'). For each combination of donor and collection batch, we generated a separate neighborhood graph based on the *scVI* embedding space (see dimensionality reduction, clustering, and visualization) and applied Leiden clustering, as implemented in the *scanpy* package, at resolution 15. We summed raw UMI counts for all cells within a cluster to generate pseudocell expression. This left 17,913 pseudocells after aggregation, with a median of 37 cells per pseudocells and a median of 342 pseudocells per donor. We removed 609 genes with nonzero expression in fewer than 10 pseudocells, leaving 34,715 genes for pseudocell analysis.

To fit the topic model, we first filtered out genes expressed in less than 10 pseudocells. We then applied Poisson negative matrix factorization to the filtered pseudocell expression data: we used k = 10 topics, and fit the model first using 400 expectation maximization (EM) updates followed by 200 stochastic coordinate descent (SCD) updates.[43] We then use the parameter estimates from this Poisson non-negative matrix factorization to recover parameter estimates for the multinomial topic model using *FastTopics*'s poisson2multinom command.

We conducted grade of membership differential expression (DE) analysis using *FastTopics*'s de_analysis function, using all expressed genes as the background set. Genes with posterior log fold change values over 2 were considered for gene set enrichment

using the Gene Ontology.[112] We performed a hypergeometric test for topic driver gene enrichment in all GO biological processes. We controlled the false discovery rate using the Benjamini-Hochberg procedure with a stringent level of 0.01, which still led to enrichment among dozens of gene sets for most topics. We focused on overrepresentation of topic DE genes in a gene set for the purposes of interpretation (OR $\geq$ 1, Table S11). Topics 3 and 9 displayed no enrichment for any GO biological processes nor clear enrichment in any of the previously characterized cell types, leading us to omit these two topics from downstream QTL calling analysis to instead focus on the 8 more clearly interpretable latent topics.

Topic eQTL calling

CellRegMap takes as inputs a kinship matrix to account for genetic similarity between cells, an environmental covariance matrix to account for similarity due to cellular context, and covariates. We used *Plink*[63] to construct the kinship matrix from all 53 donors. To construct the environmental covariance matrix, we applied an inverse normal transformation to each column (topic) of the topic loadings matrix (C, pseudocells x topics), then did the same transformation for each row (pseudocell), before generating the environmental covariance matrix (CC^T). We included sex and collection date as covariates. After QTL calling, we applied Bonferroni correction and generated a list of significant topic eQTLs as in both previous QTL analyses: a genome-wide adjusted p value threshold was defined as the Bonferroni-adjusted p value of the gene closest to a global q value threshold of 0.1, and all variants with an adjusted p value below this threshold were included in the final list (Table S12).

We evaluated the sensitivity of topic discovery and topic eQTL calling to several hyperparameters used in this study. First, we assessed the impact of the pseudocell clustering resolution. We aggregated cells into pseudocells for a range of clustering resolutions (Leiden resolution parameter = 5, 10, 15 [original resolution], 20; corresponding to 6444, 12258, 17913 [original count], and 23288 pseudocells, respectively). Then, we applied FastTopics to each pseudocell expression matrix as described, followed once again by grade of membership differential expression analysis. For each topic, this gives us a vector of log fold-changes for each gene compared to all other topics. We found a one-to-one mapping between topics across clustering resolutions by measuring the cosine similarities of these topic-specific LFC vectors (Figure S5). Topic interaction QTL calling across this range of pseudocell resolutions consistently replicated the originally reported interaction effects, as measured by the $\hat{\pi}_1$ replication rate ($\geq$ 0.92 across resolutions).

In choosing the number of topics, we aimed to balance interpretability with expressivity. We found that decreasing the number of topics from 10 to 5 reduced expressivity by merging topics corresponding to clearly distinct gene programs, such as those involved in endoderm and mesoderm differentiation, as indicated by the similar loadings of endoderm- and mesoderm-derived cell types on topic 2 in the 5-topic model (Figure S7). Increasing the number of topics to 20, however, led to more topics lacking clearly interpretable loading patterns across cell types and without significant gene set enrichment analysis results. The topic interaction eQTLs discovered with 10 topics were consistently replicated when using 5 or 20 topics, as measured by the $\hat{\pi}_1$ replication rate (0.87 for the 5 topic model, 0.9 for the 20 topic model).

Single-cell disease relevance scoring

We performed single-cell disease relevance scoring across the full single-cell HDC eQTL datasets for 40 traits (see Figure 5A), using MAGMA gene sets (top 1,000 genes) previously compiled by Zhang et al..[37] We used 500 control gene sets to compute single-cell scores. To conduct the cell type disease relevance analysis, we used Cell-ID cell type labels and the neighborhood graph based on the scVI embedding (see cell type annotation, dimensionality reduction, clustering, and visualization sections for more details).

Trait-associated cell type eQTLs

To identify schizophrenia risk variants with novel regulatory interactions identified in HDCs, we filtered the aggregate list of significant variant-gene pairs from all cell types to eQTLs that did not overlap an eQTL in any tissue according to the GTEx v8 Catalog. We then intersected these variants with genome-wide associated risk variants ($p \leq 5e{-}8$) from Trubetskoy et al.[46] For this set of genome-wide significant HDC eQTLs without GTEx overlap, we compiled a list of all variants within 1Mb of the SNP that was in LD at R2 $\geq$ 0.5 (using the 53-donor cohort in this study to compute LD). We then additionally checked these tag variants for overlap with any GTEx eQTLs in any tissue, and removed any such GTEx-overlapping tag variants to obtain our final set of schizophrenia-associated HDC eQTLs with no GTEx overlap. We used the same procedure to identify diastolic blood pressure[47] and LDL cholesterol-associated[48] HDC eQTLs with no GTEx overlap.

Phenome-wide effects of interaction eQTLs

To identify phenotypic effects of interaction eQTLs, we used an approach similar to that described for schizophrenia overlap. Instead of cell type eQTLs, we used the aggregate set of significant variant-gene pairs from all three trajectory dynamic eQTL analyses, as well as the significant variant-gene pairs from the CellRegMap analysis. We subset to interaction eQTLs that did not overlap a GTEx eQTL in any tissue, and queried the OpenTargets database for any GWAS studies in which these interaction eQTLs displayed genome-wide significant effects. We similarly removed any tagged variants with known regulatory effects in the GTEx Catalog to obtain a final set of GWAS-overlapping HDC eQTLs with no GTEx overlap.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses and visualization were performed using software listed in the key resources table. Description of statistical tests performed, with sample size and significance threshold descriptions, can be found in corresponding sections of the main text, with details described in the method details.