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

Katherine Rhodes*,1, Kenneth A Barr*1, Joshua M Popp2, Benjamin J Strober2, Alexis Battle2,3**, Yoav Gilad1**

Affiliations: 1. Department of Medicine, University of Chicago, Chicago, IL 60637, USA. 2. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA. 3. Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA.

*These authors contributed equally to this work.
**Co-corresponding authors
Abstract

Most disease-associated loci, though located in putatively regulatory regions, have not yet been confirmed to affect gene expression. One reason for this could be that we have not examined gene expression in the most relevant cell types or conditions. Indeed, even large-scale efforts to study gene expression broadly across tissues are limited by the necessity of obtaining human samples post-mortem, and almost exclusively from adults. Thus, there is an acute need to expand gene regulatory studies in humans to the most relevant cell types, tissues, and states. We propose that embryoid bodies (EBs), which are organoids that contain a multitude of cell types in dynamic states, can provide an answer. Single cell RNA-sequencing now provides a way to interrogate developmental trajectories in EBs and enhance the potential to uncover dynamic regulatory processes that would be missed in studies of static adult tissue. Here, we examined the properties of the EB model for the purpose of mapping inter-individual regulatory differences in a large variety of cell types.

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with human traits and diseases, many of which are located in noncoding regions of the genome and are putatively regulatory in function (Albert and Kruglyak 2015). Gene regulation is dynamic; it varies across cell types and conditions, including environmental exposures and cellular processes such as differentiation (Moyerbrailean et al. 2016; Strober et al. 2019; Aguet et al. 2017). To understand regulatory and functional effects of trait-associated variants, it is necessary to perform molecular assays in the relevant cell types at the relevant stages of development, and potentially to also model different environmental exposures (Umans, Battle, and Gilad 2020).

To date, there have been large scale efforts to map genetic variants that regulate gene expression (expression quantitative trait loci, or eQTLs) across human tissues. The GTEx consortium, for example, identified over 4.2 million eQTLs in 54 tissues and cell types (“The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues” 2020). Despite this enormous effort, only 11% of disease heritability can be attributed to GTEx cis-eQTLs (Yao et al. 2020), and only 43% of GWAS variants can be classified as eQTLs (Aguet et al. 2017). Thus, while a meaningful fraction of genetic disease risk can be attributed to effects on gene regulation, the majority of GWAS variants remain unexplained. This gap can be partially attributed to the fact that most studies of human gene regulation are limited to mature, adult tissues that were collected at a single point in time (and usually post-mortem). It is possible that dynamic and variable regulatory genetic effects, including those that are specific to a given cell type, time point, or environment, may underlie the mechanisms for many unexplained phenotypic associations. For example, recent efforts to characterize gene regulatory dynamics in human induced pluripotent stem cells (iPSCs) and their derived cell types have identified dynamic eQTLs that are associated with disease risk, supporting the intuitive notion that changes in gene regulation during development may play a significant role in shaping human adult phenotypes, including disease (Strober et al. 2019; Cuomo et al. 2020).
Human iPSCs offer an ethical and experimentally tractable system for the study of human development and can be differentiated into a wide variety of human cell types. Some directed differentiation protocols result in a single, homogeneous cell type, while others produce heterogeneous populations of cells. In particular, iPSCs can be used to form embryoid bodies (EBs). EBs are three dimensional aggregates of spontaneously and asynchronously differentiating cells; that is, they contain developmentally diverse cell types from all three germ layers. EB formation has been used to verify stem cell pluripotency for decades; yet, until recently, the complexity of EB cellular composition has precluded their use in genomic studies. With single-cell RNA-sequencing (scRNA-seq), it is now possible to characterize the numerous spatially and developmentally distinct cell types within EBs, including transient cell types that would otherwise be inaccessible. Indeed, recent scRNA-seq studies of human EB differentiation have revealed the diversity of cell types composing these structures and the transcriptional dynamics governing early fate decisions (Han et al. 2018; Guo et al. 2019).

The application of scRNA-seq to EBs should make it possible to map eQTLs in a multitude of cell, including developmental and transient cell types that are no longer found in adult tissues. The ability to grow multiple cell types in the same dish obviates the need for multiple directed differentiation protocols and affords greater control over confounding variables that might mask genetic effects on gene expression. To date, however, the only studies that have sequenced EB cells have relied on a small sample of cells from a single individual, leaving a gap in our understanding of technical, biological, and inter-individual variation present in this system (Han et al. 2018; Guo et al. 2019).

Understanding the sources of variation that affect scRNA-seq data from EBs is crucial for evaluating the utility of EBs as a novel system for eQTL discovery. To this end, we generated EBs from three individuals in three replicates. We do not report here an eQTL mapping study, but rather the first pilot study to examine the utility of this potentially transformative in vitro system. Our study design allowed us to explore the biological and technical variation present at all levels of the EB data set; specifically, we evaluated the consistency in cell type composition across replicates and individuals, characterized the structure of variation in gene expression across the entire data set, and finally, captured patterns of dynamic gene expression along distinct developmental trajectories. Our results indicate that scRNA-seq of differentiating EBs has the potential to be a powerful model system for the study of inter-individual variation in gene regulation across an array of functionally and temporally diverse cell types.

Results

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

**Study design, data collection, and preprocessing**

To characterize sources of variation in gene expression in human EBs, we differentiated EBs from three human iPSC lines in three replicates (see Materials and Methods). We performed the experiment in three batches, where each batch includes one replicate from each of the three individuals. EBs differentiate quickly, with cell types representing endoderm, mesoderm, and ectoderm present after eight days (Han et al 2018). In this study, we maintained EBs for three weeks after formation, allowing cells to continue to differentiate and mature. After 21 days, we collected scRNA-seq data, targeting equal numbers of cells from each individual in each replicate. After filtering and quality control (Materials and Methods), we retained high-quality data from a sample of 42,488 cells (an average of 4,721 cells per individual/replicate). For these cells, we obtained a median of 16,712 UMI counts per cell, which allowed us to measure the expression of a median of 4,274 genes per cell (Fig. S1). We integrated data from all cells using Harmony, which anchors the data sets by cell type (Korsunsky et al. 2019).

**Cell type composition**

To validate our expectation that EBs should contain cells from each germ layer, we first characterized the expression of early developmental marker genes. We found cells expressing markers for endoderm (SOX17, FOXA2), mesoderm (HAND1), and ectoderm (PAX6), in addition to cells still expressing pluripotency markers (POU5F1, MYC, NANOG). We visualized the data with uniform manifold approximation and projection (UMAP) and observed that cells expressing each of these germ layer markers occupied distinct groups in UMAP space (Fig. 1A-D) (Becht et al. 2019). Moreover, we found that every replicate in our experiment, regardless of the individual, includes cells from all three germ layers (Fig. 1E, Fig. 1G).

We next sought to further explore the heterogeneous cell types present in these EBs. In studies of scRNA-seq data from tissues and samples with well-characterized cell type composition, clustering is often applied to demarcate populations of pure cell types within heterogeneous samples. In these studies, clustering resolution, which determines the number of clusters identified by the algorithm, is typically chosen to recapitulate the expected number of known cell types. The identified clusters can be annotated based on the expression of known marker genes.

In our case, however, we had no *a priori* knowledge of the exact number or types of cells that would result from the spontaneous differentiation of the EBs. Hence, we used three complementary approaches to annotate cells, capturing various perspectives on what might define a cell type in this data set. First, we identified cell types by clustering cells and annotating the cell types based on the genes that are highly expressed in each cluster. Second, we annotated cell types by considering the correlation of gene expression in our data with a reference
data set of known primary cell types. For our third approach we used a different perspective, and applied topic modeling to consider a less discrete definition of cell type.

For the first approach we used a standard clustering analysis, the Louvain algorithm in Seurat, to identify groups of cells with similar transcriptomes (Blondel et al. 2008). To avoid making assumptions about the true number of cell types present, we repeated this analysis across different clustering resolutions (resolution 0.1, 0.5, 0.8, and 1). As expected, the number of clusters we identified varied greatly depending on the resolution (Fig.1E-F, Fig. S2). We performed each subsequent analysis using clusters defined at multiple resolutions, to ensure that our qualitative conclusions are robust with respect to the number of clusters identified.

For each clustering resolution, we calculated pseudobulk gene expression levels using cells from the same cluster, individual, and replicate. To identify marker genes expressed in each cluster, we used Limma and voom
to perform differential expression analysis (Materials and Methods) using the pseudobulk estimates. For example, considering the gene expression data of the seven clusters identified at resolution 0.1, we found that the most significantly upregulated genes in each cluster included known marker genes for pluripotent cells (cluster 0), early ectoderm (cluster 1), mesoderm (cluster 2), neural crest (cluster 3), endoderm (cluster 4), neurons (cluster 5), and endothelial cells (cluster 6) (Fig. 1E, Table 1). Using this approach provides a confident set of broad cell type categories present in these data. At higher resolutions, DE analysis between clusters enabled annotation of some more specific cell types; for example, at clustering resolution 1, cluster 19 differential expression of hepatocyte marker genes FGB, TTR, and AFP. However, confident cell type classification of Seurat clusters at higher resolution based on DE alone proved difficult (Tables S3-6).

<table>
<thead>
<tr>
<th>Cluster (Res. 0.1)</th>
<th># cells in cluster</th>
<th>Top 10 marker genes by adj. P</th>
<th>Top 10 marker genes by logFC</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17693</td>
<td>TERF1, PHC1, SEPHS1, UGP2, DPPA4, TBC1D23, JARID2, USO1, ZNF398, LRRC47</td>
<td>DPPA5, DPPA3, GDF3, NANOG, FGF4, POU5F1, CBR3, PRDM14, DPPA2, TRIML2</td>
<td>Pluripotent Cells</td>
</tr>
<tr>
<td>1</td>
<td>14383</td>
<td>TPBG, FGFBP3, FZD3, LIX1, SDK2, BTBD17, DACH1, PLAGL1, DEK, ZNF219</td>
<td>FEZF2, EMX2, LHX2, SOX3, PAX6, WNT7, BARX, SOX1, ZIC1, SIX3</td>
<td>Early Ectoderm</td>
</tr>
<tr>
<td>2</td>
<td>3086</td>
<td>TNN1, COL6A3, COL5A1, RGS4, ACTA2, TMEM88, DOK4, SLC40A1, HAND2, COL3A1</td>
<td>RGS13, LUM, TECRL, DCN, HAND1, PITX1, COL3A1, SLN, IGF2, FIBIN</td>
<td>Mesoderm</td>
</tr>
<tr>
<td>3</td>
<td>2673</td>
<td>NR2F1, CNP, S100B, EDNRA, FGFBP3, ATP1A2, DNAJC1, ZEB2, PHACTR3, METRN</td>
<td>MPZ, PRSS56, ROPN1, SOX10, S100B, SCRGR1, NPR3, MOXD1, TFAP2B, PHACTR3</td>
<td>Neural Crest</td>
</tr>
<tr>
<td>4</td>
<td>2368</td>
<td>S100A16, LGALS3, GATA3, CST3, KRT19, FN1, EPST11, DYNLT3, HDHD3, PKP2</td>
<td>APOA2, CST1, APOA1, APOC3, FGB, RBP4, S100A14, TTR, FGA, APOB</td>
<td>Endoderm</td>
</tr>
<tr>
<td>5</td>
<td>1990</td>
<td>TAGLN3, RTN1, NHLH1, STMN2, ELAVL2, FNDC5, PEBP4, ELAVL4, DCX, MLLT11</td>
<td>NEUROD1, NHLH1, STMN2, NEUROD4, TBR1, STMN4, NEUROG1, SST, ELAVL3, SLC17A6</td>
<td>Neurons</td>
</tr>
<tr>
<td>6</td>
<td>295</td>
<td>EGFL7, GNG11, RAMP2, IGFBP4, PPM1F, RASGRP3, RCSD1, MAP4K2, PLVAP, DOCK6</td>
<td>PLVAP, CD34, CD93, CDH5, DIPK2B, PECAM1, EMCN, CRHBP, ESAM, ECSCR</td>
<td>Endothelial Cells</td>
</tr>
</tbody>
</table>

Table 1: Classification of Seurat cluster identity (clustering resolution 0.1) based on differential expression of marker genes.
To pursue the second approach, we annotated cells by comparing our gene expression data to available reference sets of scRNA-seq data from primary fetal tissues, human embryonic stem cells (hESCs), and hESC-derived EBs (Cao et al. 2020; Han et al. 2020). To do this, we first integrated our data set with the reference data sets and visualized cells with UMAP (Fig. 2A, Fig. S3, Materials and Methods, Supplementary Materials and Methods). We observed that reference hESCs cluster closely with pluripotent EB cells. We also observed that the hESC-derived EBs and our iPSC-derived EBs tend to occupy the same areas in UMAP space, implying high overall similarity in cell type composition despite differing protocols for EB differentiation (and despite the fact that the experiments were performed in different labs). EB cells also show overlap with many primary fetal cell types (Fig. 2B-C, Fig. S4). For example, EB cells annotated as neural crest based on our gene expression analysis, overlap with primary fetal cell types derived from neural crest, like Schwann cells and enteric nervous system (ENS) glia (Fig. 2B). EB cells annotated as neurons based on our gene expression analysis overlap with fetal neuronal subtypes, including inhibitory neurons, excitatory neurons, granule neurons, ENS neurons, and others. EB cells also show overlap with populations of cells that are rare in the fetal data set, including AFP_ALB positive cells (hepatic cells), thymic epithelial cells, and lens fibre cells (Fig. 2C).

Encouraged by these observations, we expanded the annotation of our EB cells (which up to this point were based on the expression of known marker genes) by using the known annotations of the reference primary fetal cell type data set. Specifically, we transferred cell annotations to EB cells based on the nearest neighbor reference cell (Fig. 2D). Using this approach, we found EB cells representing 66 of the 77 primary cell types present in the reference fetal data set (Table S1). The most common annotation was hESC; this can be partially attributed to the high proportion of pluripotent cells in our EB data set, but also to the fact that the reference fetal data set does not include many early developmental cell types. Indeed, many cells annotated as hESC here are likely to represent immature, differentiating cells which are no longer pluripotent but whose transcriptional profiles more closely match hESCs than the more highly differentiated fetal cell types present in the reference data set. In this sense, EB data sets may capture transient developmental cell types that are difficult or impossible to study even in fetal primary samples. Outside the hESCs, many fetal cell types are only represented by small populations of EB cells. For example, only one EB cell is annotated as a sympathoblast, and only two cells are annotated as thymocytes. These observations indicate that we need a deeper sampling of EBs in order to properly explore their true cell type composition. Overall, annotation based on the reference set revealed the presence of dozens of diverse cell types in EBs.

**Topic modeling of the single cell gene expression data**

Both of the approaches we described above (clustering, and comparison to a reference data set) assume that “cell types” are discrete categories. Accordingly, each cell has a single true identity, and the cell type categories are assumed to be homogeneous and static. This definition of a cell type is intuitive and makes it practical to consider results from single cell analysis in the context of the wealth of knowledge previously gained from bulk
Figure 2: Reference Integration and cell type annotation. A) UMAP visualization of EB cells from this study and cells from reference data sets of fetal cell types, Day 20 EBs, and hESCs after integration. Cells are colored by data set. B) UMAP visualization of EB cells from this study and cells from the fetal reference after integration. Cells are colored by Seurat cluster identity at clustering resolution 0.1, with grey points representing cells from the fetal reference set. C) UMAP visualization of EB cells from this study and data from the fetal reference after integration. Cells are colored by cell types present in the fetal reference data set, with grey points representing EB cells. D) UMAP visualization of EB cells from this data set with annotations transferred from the fetal and hESC reference sets.
assays. However, partitioning cells into discrete groups is unlikely to capture the full degree of heterogeneity in gene expression of single cells. A particular cluster or cell “type” may collapse multiple cell states, obscuring functionally distinct subgroups such as cells in different stages of the cell cycle. This problem becomes more apparent in data sets that include differentiating cells, which are expected to show varying degrees of similarity to a terminal cell type. In an alternate paradigm, cell type can be viewed as continuous, with the expression profile of each cell representing grades of membership to multiple categories (Dey, Hsiao, and Stephens 2017). One method used to capture cell identity in this paradigm is topic modeling, which learns major patterns in gene expression within the data set, or topics, and models each cell as a combination of these topics.

We applied topic modeling using fastTopics at a range of topic resolutions, identifying 6, 10, 15, 25, and 30 topics in our data. Some topics correspond closely to Seurat clusters, showing loading on cells of a given cluster but not on others (Fig S7). For example, in the k=6 topic analysis, topic 1 is loaded exclusively on cells assigned to

Figure 3: Topic modelling of EB cells. A) Structure plot showing the results of topic modelling at k=6. Plot includes a random subset of 5000 EB cells divided by seurat cluster at resolution 0.1. B) UMAP projection of cells colored by loading of topic 1. C) Box plot showing the loading of topic 1 from the k=6 topic analysis on each Seurat cluster at clustering resolution 0.1. D) Box plot showing the loading of topic 1 from the k=6 topic analysis on each Seurat cluster at clustering resolution 1. E) Volcano Plot showing genes differentially expressed between topic 1 and all other topics from the k=6 topic analysis. Points are colored by the average count on the logarithmic scale.
Seurat cluster 4 (cluster resolution 0.1) which we previously annotated as endoderm (Fig. 3A-D, Fig. 1E, Table 1). Compared to other topics, topic 1 shows significant increase in expression of $FN1$ and $AFP$, which are known markers of hepatocytes (Fig. 3E, Table 2). Seurat clustering at higher resolution (resolution 1) results in further categorical division of this large endoderm group of cells into definitive endoderm and hepatocytes (Fig. 1F). Topic modeling revealed that these cells actually exhibit variable grades of membership in topic 1 (in k=6 topic model); this gradient captures a temporal continuum of differentiation.

Certain topics are shared across cells assigned to different Seurat clusters (Fig. 3A, Fig. S7). For example, topic 6 from the k=6 topic analysis is loaded across all Seurat clusters; compared to all other topics, topic 6 shows increased expression of many ribosomal genes, housekeeping genes ($GAPDH$), and genes coding for proteins involved in cellular metabolism ($LDHA$) (Fig. S5-7). This indicates that topic 6 captures patterns of gene expression associated with cellular processes and functions that are not specific to a particular cell type. This again highlights an advantage of topic modeling, enabling us to explore variation in the representation of gene expression profiles associated with processes shared across many cell types, simultaneously with identifying cell-type specific patterns.

### Biological and technical sources of variation

Once we functionally annotated EB cells using the three approaches discussed above, we sought to understand the consistency in cell type composition across individuals and between replicates. Here, “replicate” corresponds to a batch of EB differentiations in which each cell line was differentiated, dissociated, and collected in tandem; “replicate” therefore captures technical variation related to differentiation batch, dissociation batch, and single cell collection batch. We began by calculating the proportion of cells that were assigned to each Seurat cluster at resolution 0.1 for each replicate. We then performed hierarchical clustering of the samples based on the proportion of cells in each Seurat cluster (Fig. S8). Using this approach, replicate-individual samples cluster first by individual, indicating that cell type composition is distinct between individuals and is consistent between replicates of each individual. We repeated this analysis at a range of cluster resolutions and determined that this finding is robust with respect to the number of clusters (Fig. S8).

### Table 1: Top 15 driver genes of each topic from the k=6 topic model based on Z-score.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Top 15 driver genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>k1</td>
<td>S100A10, FTL, FN1, APOC1, CST3, APOE, SERPINE2, KRT19, CKB, S100A11, LGALS3, TMSB10, S100A16, AFP, PTGR1</td>
</tr>
<tr>
<td>k2</td>
<td>MT-CO2, MT-CO3, MT-CO1, MT-CYB, PRDX1, MT-ND4, MT-ATP6, GSTP1, MT-ND1, RPL8, APOE, RPSA, RPL12, PFN1, HMGA1</td>
</tr>
<tr>
<td>k3</td>
<td>PTMA, NCL, RPL23, SET, HSP90AB1, TPL27A, MT-ND4, L1TD1, SERBP1, TERF1, HSDP1, CENPF, DPPA4, MT-ATP6, UGP2</td>
</tr>
<tr>
<td>k4</td>
<td>S100A10, KRT19, S100A11, VIM, MDK, TMSB10, KRT8, SPARC, COL1A1, FN1, COL1A2, COL6A2, KRT18, TPM1, ANXA2</td>
</tr>
<tr>
<td>k5</td>
<td>TUBA1A, VIM, MARCKSL1, MARCKS, TUBA1B, MAP1B, ID3, CRABP1, PTMS, TMSB10, H1FX, STMN1, CENPV, CRABP2, NUCKS1</td>
</tr>
<tr>
<td>k6</td>
<td>RPS27, VIM, LDHA, GAPDH, IGFBP2, TUBA1A, APOA1, RPL13, TMSB10, S100A10, RPL6, RPL30, RPL9, RPS19, RPL37</td>
</tr>
</tbody>
</table>
We also repeated this analysis using topic loadings as a measure of cell type composition. We calculated the loading of each topic on each individual-replicate group and performed hierarchical clustering (Fig. S9). Again, we found that at varying values of k, samples generally cluster by individual, but using the higher resolution topic-based approach, we also observed substantial variation between replicates (Fig. S9). One individual, 18858, always clusters away from the other two lines, showing a distinct distribution of cell types that is consistent across all replicates. The EBs from this individual have a particularly high proportion of pluripotent cells and a lower differentiation efficiency compared to the other two lines (Fig. 1G).

We further characterized determinants of variation in our system by considering factors that contribute to variation in gene expression levels. Hierarchical clustering of pseudobulk expression estimates of cells from the sample Seurat cluster (res. 0.1), replicate, and individual shows that, as might be expected, samples tend to cluster first by cell type (Seurat cluster), then by individual and replicate (Fig. 4A). We performed variance partitioning using pseudobulk expression levels to estimate the relative contribution of cell type, individual, and replicate to overall patterns of gene expression variation (Fig. 4B) (Hoffman and Schadt 2016). We found that replicate and individual explained approximately equal proportions of the variance (each explains a median value of ~5% of variance). Cell type identity explained the largest proportion of variation at all clustering resolutions tested (variance explained median value ~60% at clustering resolution 0.1), although this figure is likely exaggerated since cell type identity is determined by clustering, which will inherently maximize variation between cell types. Depending on clustering resolution, a median value of approximately 20-30% of variance is explained by residuals, which can be attributed to noise or other technical variation not specifically modeled (Fig. S10). We then partitioned the variance in gene expression at single cell resolution (instead of using pseudobulk estimates) and found that replicate explains more variation on average than individuals, with cell type identity continuing to explain more variance than either (Fig. 4C). At single cell resolution, residuals explain a median value of 93% of the variation, which is expected due to the high degree of variance (both biological and technical) in gene expression profiles among individual cells.

To determine whether biological and technical factors contributed differently to variation between cell types, we also partitioned the variance due to replicate and individual in each Seurat cluster separately (Fig. S11-12). The results are not uniform across clusters. At clustering resolution 0.1, individual contributes more to variation, on average, in clusters 0, 1, 4, and 5, while replicate contributes more to variation in clusters 2, 3, and 6. Notably, clusters 2, 3, and 6 include only a few cells from individual 18858 (Table S2). Studies that incorporate a larger number of cells will increase representation of rare cell types, which will increase power to study patterns of gene regulation. In every cluster, variation due to replicate dominates the variation of certain genes but not others. This complex structure indicates that, unlike most other eQTL studies – where adding individuals is always
preferable to adding technical replicates - future studies of EBs need to implement study designs with multiple replicates to appropriately account for this variation.

Figure 4: Exploration of the biological and technical variation in gene expression across EB cells. A) Heatmap showing hierarchical clustering of cells based on normalized gene expression. This analysis uses only gene expressed in at least 20% of cells in at least one cluster (at clustering resolution 0.1) and does not include ribosomal genes. B) Violin plot showing the percent of variance in gene expression explained by cluster (resolution 0.1), replicate, and individual in this data set after partitioning variance in pseudobulk samples. C) Violin plot showing the percent of variance in gene expression explained by cluster (resolution 0.1), replicate, and individual in this data set after partitioning variance at single cell resolution.
Because individual variation contributes significantly to overall patterns of variation in gene expression, EBs have the potential to be a powerful model to understand inter-individual variation in gene regulation across cell types and to map dynamic eQTLs. We performed a power analysis to better understand the relationship between power, sample size, and the total number of individual cells analyzed, or the experiment size (Fig. 5). Assuming a simple linear regression to map eQTLs and a conservative Bonferroni correction for multiple testing (FWER = 0.05, Methods), we estimated that an experiment consisting of 58 individuals with 3,000 cells collected per individual, collected across 3 replicates (experiment size of 174,000 cells total), would provide 93% power to detect eQTLs with a standardized effect size of 0.6. These assumptions represent an experimentally tractable study design, and a conservative estimate of standard and dynamic eQTL effect sizes, suggesting this could be a powerful system for QTL studies across diverse human cell types.

![Figure 5: Power to detect eQTLs.](image-url)

**Figure 5: Power to detect eQTLs.** Power is a function of effect size, sample size, experiment size, and significance level. Power curves are computed for a range of sample sizes and experiment sizes (cells per individual).

**Dynamic patterns of gene expression**

Arguably, the most attractive property of single cell data from the EB system is the ability to study dynamic gene regulatory patterns throughout differentiation. In order to explore dynamic patterns of gene expression, we inferred developmental trajectories using PAGA (Wolf et al. 2019). The coarse-grained PAGA graph shows edges that represent likely connections between cell clusters (clustering resolution 1) and we were able to trace developmental trajectories through this path (Fig. 6A-C). Since the EBs still include undifferentiated pluripotent cells, we were able to define rooted trajectories beginning at the known starting point. Using this approach, we inferred differentiation branches to each germ layer originating from cluster 22, which expresses primitive streak markers (*MIXL1*, *EOMES*), showing recapitulation of developmental trajectories defined during gastrulation (Fig 6B-C, Fig. 1F, Fig. S13). We also note that hepatocytes (cluster 19), an endoderm-derived cell type, branch off of the endoderm cluster (cluster 10). Endothelial cells (cluster 24), which are derived from mesoderm, branch off from the mesoderm cluster (cluster 4), and neurons (clusters 12, 15), an ectoderm-derived cell type, branch off from the early ectoderm clusters (clusters 2, 3, 8, 9, 13, 14, 17, 20, 21, 26, and 27) (Fig. 6A-C, Fig. 1F, Table1).
We then assigned pseudo-time values to each cell using the diffusion pseudo-time method with pluripotent cells (cluster 1) defined as the root (Haghverdi et al. 2016) (Fig. 6B-C, Fig. S14, Fig. 1F). We manually traced high confidence trajectories through the data representing the progression from pluripotent cells to hepatocytes (clusters 0, 1, 5, 6, 7, 10, 11, 16, 18, 19, 25, and 22), pluripotent cells to endothelial cells (clusters 0, 1, 4, 5, 6, 7, 11, 16, 18, 22, 24, and 25), and pluripotent cells to neurons (clusters 0, 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 15, 26, and 27) (Fig. 6A-C). For groups of clusters with a higher degree of connectivity (e.g. clusters expressing pluripotent markers and clusters expressing early ectoderm markers), all clusters within the region with high connectivity were included in the trajectory to avoid choosing an arbitrary path through these clusters. Next, we applied split-GPM, an unsupervised probabilistic model, to infer dynamic patterns of gene expression within a particular developmental trajectory, while simultaneously performing clustering of genes and samples. Split-GPM is built for use with time course, bulk RNA-seq data; therefore, we calculated pseudobulk expression values for individual-replicate groups within decile bins of pseudo-time. We were able to identify gene modules with distinct dynamic patterns of expression along the trajectories to neurons, hepatocytes, and endothelial cells (Fig. S15-17).

Gene set enrichment analysis of these modules shows expected dynamic patterns (Fig. S15-17). For example, we found that gene modules that increase expression through pseudo-time along the differentiation trajectory to hepatocytes, which are the predominant cell type of the liver and are responsible for the production of bile, are enriched for the hallmark bile acid metabolism and fatty acid metabolism gene sets. In the trajectory leading to endothelial cells, which are derived from mesoderm, we found that a gene module with high expression at intermediate pseudo-time values is enriched for hallmark genes expressed during the epithelial-mesenchymal transition, which is essential for mesoderm formation (Evseenko et al. 2010). In all three trajectories, gene modules characterized with higher expression at low pseudo-time values show enrichment for gene sets related to the cell cycle; this is expected because pluripotent cells at the lowest pseudo-time values tend to grow and divide faster than more differentiated and more mature cell types, which often exit the cell cycle (Buttitta and Edgar 2007).

To determine the consistency in dynamic patterns of gene expression between replicates and individuals, we ran split-GPM ten times on cells from the neuron, hepatocyte, and endothelial cell lineages and observed how often each pair of individual-replicate samples clustered together (Fig. 6D-F). In the neuronal and endothelial lineages, all three replicates of 18511 always clustered together and often cluster with replicates of 19160, indicating that these two lines share similar expression dynamics in these trajectories (Fig. 6D, F). Replicates of 18858 often cluster together and rarely clustered with the other individuals, suggesting that not only did 18858 have poor differentiation efficiency, but cells that did differentiate show a distinct pattern of expression dynamics. In the hepatocyte lineage, we observed stronger replicate-specific differences (Fig. 6E). Replicates of individual
19160 still tended to cluster together and to cluster with replicates 1 and 2 of 18511. Replicate 3 of 18511 rarely clustered with the other replicates of that individual, indicating that there were replicate-specific effects on dynamic gene expression.

**Figure 6: Trajectory inference and identification of dynamic gene modules.** A) Force atlas plot of EB cells colored by broad cell type categories corresponding to Seurat clusters identified at clustering resolution 0.1 (see Table 1). B) Force atlas plot of EB cells colored by Seurat cluster at clustering resolution 1. C) Coarse-grained PAGA graph showing inferred edges between Seurat clusters defined at clustering resolution 1. D-F) Heatmaps showing the frequency with which individual-replicate groups were assigned to the same cluster after running split-GPM 10 times in the neuronal, hepatic, and endothelial lineages.

**Discussion**

This work represents a thorough exploration of heterogeneity in single cell data obtained from human EBs towards the goal of establishing this system as a tool to enable studies of variation in human gene regulation across a range of spatially and temporally diverse cell types. We used iPSC-derived embryoid bodies because this *in vitro* model system circumvents the logistical challenges and ethical barriers associated with studies of primary human developmental tissues. This system has key advantages over studies of primary tissues; for example, we are able to control cellular environment in vitro and intentionally design experiments with respect to biological factors including age, sex, and ancestry. Further, we can generate EBs comprised of the same set of diverse cell types from large samples of individuals, enabling high-powered comparisons of cell-type specific gene expression.
In subsequent studies we plan to leverage embryoid bodies to identify QTLs and dynamic QTLs across diverse terminal and differentiating cell types. This, of course, raises an ostensibly critical question: to what extent do the cell types derived from embryoid bodies faithfully model immature, developing cells \textit{in vivo}? There is no doubt that the \textit{in vitro} EB differentiated cells are not a perfect model of primary cell types. The question is whether EB cells are sufficiently representative of primary cell types to be informative. To address this question, we performed several analyses, which suggest that the EB model can be useful. Specifically, we found that EB cells express known cell-type specific marker genes, including markers of known developmental stages. EB cells also cluster with more than 60 diverse primary cell types from a reference panel of fetal tissues and hESCs, including rare fetal cell types. Lastly, we identified gene modules with dynamic expression patterns that match broad expectations of developmental biology. Together, these results provide evidence that EBs are a suitable model of both terminal and developmental cell types.

Moreover, EBs may be a useful model for understanding the genetic underpinnings of human traits and diseases regardless of the degree to which they faithfully model human development. EB-derived cells represent a wealth of previously unstudied cell states and dynamic processes. Hypothetically, QTLs identified in these cell types still represent biologically meaningful differences in genetic control of gene regulation, whether they manifest in human development or in adult tissues upon a particular environmental exposure. To provide an anecdotal example of this reasoning, we considered previously collected data from an \textit{in vitro} differentiation experiment. We took a closer look at the 28 middle-dynamic eQTLs Strober \textit{et al.} identified during the differentiation of iPSCs to cardiomyocytes (Strober \textit{et al.} 2019). Middle-dynamic eQTLs have their strongest genetic effect at intermediate stages of the differentiation time course, and most of them (25/28) were identified exclusively at these intermediate stages of differentiation. This means that these eQTLs are active in early \textit{in vitro} differentiating cells whose fidelity to primary developing cell types has not been ascertained. These 28 dynamic eQTLs were entirely novel and had not been identified as \textit{cis} eQTLs in any tissue in the GTEx data set. Strober \textit{et al.} reported that one of these middle-dynamic eQTLs was also found to overlap a GWAS variant associated with body mass index and red blood cell count. This finding highlights that dynamic eQTLs acting in early cell types in \textit{in vitro} differentiations may affect long-term disease risk in adults.

To further explore the utility of dynamic eQTLs identified in \textit{in vitro} differentiations, we used GTEx data to ask whether the middle-dynamic eQTLs are associated with inter-individual variation in \textit{trans} gene expression or cell composition, either of which could indicate lasting downstream effects in adult tissue from transient dynamic \textit{cis} eQTLs. \textit{Trans} eQTL associations are more tissue-specific than \textit{cis} eQTLs, but \textit{trans} eQTLs are much harder to identify because of their small effect sizes and the requirement to test the association of every locus with every gene. Here, we identified a middle dynamic eQTL SNP (rs6700162) from Strober \textit{et al.} that, in GTEx data, is associated with fibroblast cell type proportions in HLV (heart left ventricle; \(p < 0.0009\)) and with cardiac muscle cell proportions in HLV (\(p < 0.003\)). This SNP was also found to have a \textit{trans} eQTL p-value of \(1 \times 10^{-5}\) in coronary
artery. Without the prior knowledge provided by dynamic eQTL data from the *in vitro* differentiated cardiomyocytes, it would have been impossible to identify these associations using adult primary tissues because the burden of multiple testing within the entire GTEx data set is considered is prohibitively large. This example implies that developing EB cells could be used to understand how transient effects on gene expression are propagated into functional, long-lasting consequences downstream.

In summary, human embryoid bodies have the potential to be a powerful system for the identification of dynamic eQTLs. In this pilot study, we performed foundational analyses to better understand how to appropriately conceptualize heterogeneity in this kind of data and how to best design large-scale studies of embryoid bodies. We explored cell type composition of embryoid bodies in three paradigms; first, with discrete cell types identified with a traditional clustering algorithm, then with more continuous cell “types” identified with topic modeling, and finally exploring dynamic changes in gene expression along trajectories using pseudotime. Cell types defined by discrete clustering are often easier to interpret because they can be contextualized with marker genes and reference data sets defined with bulk sequencing. We conclude, however, that topic modeling is more appropriate for highly heterogeneous single cell data sets like this one. We also explored sources of variation in cell type composition and gene expression. We found that individual variation primarily contributes to patterns in cell type composition based on both discrete clustering and topic modeling. However, variation between replicates is non-negligible, indicating that future studies should focus on inter-individual variation in cell type composition. We also found that technical variation between replicates contributes significantly to variation in gene expression. Future efforts to map regulatory QTLs in EBs should implement study designs with multiple replicates to appropriately correct for batch effects. Overall, this pilot study has laid the groundwork to transform embryoid bodies into a powerful model system for the understanding of human gene regulation.

### Materials and Methods

#### Samples

We used iPSC lines from three unrelated individuals from the Yoruba HapMap population to form EBs. The iPSC lines were reprogrammed from lymphoblastoid cell lines and were characterized previously (Banovich et al. 2016). Two of the lines (18511, 18858) are from female individuals and one (19160) is from a male individual.

#### iPSC maintenance

We maintained feeder-free iPSC cultures on Matrigel Growth Factor Reduced Matrix (CB-40230, Thermo Fisher Scientific) with StemFlex Medium (A3349401, Thermo Fisher Scientific) and Penicillin/Streptomycin (30002Cl, Corning). We grew cells in an incubator at 37°C, 5% CO2, and atmospheric O2. Every 3-5 days thereafter, we passaged cells to a new dish using a dissociation reagent (0.5 mM EDTA, 300 mM NaCl in PBS) and seeded cells with ROCK inhibitor Y-27632 (ab120129, Abcam).
Embryoid body formation and maintenance
We formed EBs using a modified version of the STEMCELL Aggrewell400 protocol. Briefly, we coated wells of an Aggrewell 400 24-well plate (34415, STEMCELL) with anti-adherence rinsing solution (07010, STEMCELL). We dissociated iPSCs and seeded them into the Aggrewell400 24-well plate at a density of 1,000 cells per microwell (1.2x10^6 cells per well) in Aggrewell EB Formation Medium (05893, STEMCELL). After 24 hours, we replaced half of the spent media with fresh Aggrewell EB Formation Medium. 48 hours after seeding the Aggrewell plate, we harvested EBs and moved them to an ultra-low attachment 6-well plate (CLS3471-24EA, Sigma) in E6 media (A1516401, ThermoFisher Scientific). We maintained EBs in culture for an additional 19 days, replacing media with fresh E6 every 48 hours. We performed three replicates of EB formation on different days; each replicate included all three lines.

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

Single cell sequencing
We collected scRNA-seq data using the 10X Genomics V3.0 kit. Single cell collections for this experiment were mixed with cells from a larger experiment in all three replicates. From the first replicate of EB differentiations, we mixed YRI cells with EB cells from an additional three humans and chimpanzees (nine individuals total). Even numbers of cells from all nine individuals were collected across nine lanes of a 10x chip, targeting 10,000 cells per lane. In this replicate, reagents from three different 10x kits were used. From replicates 2 and 3 of EB differentiation, EBs were only generated from the YRI individuals and the three chimpanzees (six individuals total). In each replicate, we mixed even numbers of cells of each individual and collected cells in four lanes of a 10X chip, targeting 10,000 cells per lane, and samples were processed using reagents from a single 10X kit.

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

**Alignment and sample deconvolution**
We used STARsolo to align samples to both the human genome (GRCh38) (Dobin et al. 2013) and the chimpanzee genome (January 2018; Clint_PTRv2/panTro6). We used gene annotations from ensembl98 and transmapV5 respectively. In order to separate human cells from chimpanzee cells we identified discordant reads – those that mapped with different scores in each species. We identified a cell as human if 1) at least 5 discordant reads that had a higher human mapping score and 2) at least 80% of discordant reads had a higher human mapping score. The remainder of analyses in this work were restricted to these human cells. We demultiplexed individual samples and identified doublets using demuxlet (Kang et al. 2018). For this demultiplexing with demuxlet, we used previously collected and imputed genotype data for these three Yoruba individuals from the HapMap and 1000 Genomes Project (Auton et al. 2015; Belmont et al. 2003).

**Filtering and integration**
We ran EmptyDrops to identify barcodes tagging empty droplets and kept only barcodes with a high probability of tagging a cell-containing droplet (i.e. we kept cells with an EmptyDrops FDR < 0.0001) (Lun et al. 2019). We removed cells labeled as doublets or ambiguous by demuxlet, keeping only barcodes labeled as singlets. We also filtered the data to include only high-quality cells expressing between 3-20% mitochondrial reads and expressing more than 1,500 genes. We normalized data from each 10X sequencing lane using SCTransform in Seurat (Butler et al. 2018; Hafemeister and Satija 2019). In total, we obtained 42,488 high quality cells. We then merged data from each of the 10X lanes from all replicates, scaled the data, and ran principal components analysis (PCA) using 5,000 variable features. We then integrated data with Harmony to correct the PCA embeddings for batch effects and individual effects (Korsunsky et al. 2019).

**Clustering and cell type annotation**
To cluster the data, we applied Seurat’s FindNeighbors using 100 dimensions from the Harmony-corrected reduced dimensions, followed by FindClusters at resolutions 0.1, 0.5, 0.8, and 1.
We performed differential expression analysis using the limma R package (Ritchie et al. 2015). First, we filtered genes to include only those expressed in at least 20% of cells in at least one cluster at a given clustering resolution. We then calculated pseudobulk expression values for each individual-replicate-cluster grouping (i.e. cells from the same individual, same replicate, and same cluster assignment). Accordingly, we define pseudobulk expression values as the sum of normalized counts in each group. Next we TMM-normalized pseudobulk expression values and used voom to calculate a weighted gene expression value to account for the mean-variance relationship (Law et al. 2014). We then fit the following linear model:

$$Y = 0 + \beta_{\text{cluster}} * x + \beta_{\text{replicate}} * x + \beta_{\text{individual}} * x$$

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

We next compared cells to reference data sets of primary fetal cell types, Day 20 hESC-derived EBs, and hESCs (Cao et al. 2020; Han et al. 2020). To integrate our cells with the reference sets, we first subset each reference set to include only protein coding genes. Because the Cao et al. reference set is so large, containing over 4 million cells, we subset cells from this reference set to include a maximum of 500 cells per cell type. We then normalized each reference set using SCTransform (Butler et al. 2018; Hafemeister and Satija 2019). We then merged the data sets using Seurat, re-ran SCTransform regressing out data set specific effects of sequencing depth, scaled the data, and ran PCA. We then ran Harmony to correct PCA embeddings for the effects of each data set to complete the integration (Korsunsky et al. 2019). We then transferred cell type annotations from cell types present in the fetal reference and hESC to EB cells. For each EB cell, we found the nearest reference cell in Harmony-corrected PCA space based on Euclidean distance and transferred the cell type annotation.

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

Hierarchical clustering based on cell type composition and gene expression

To understand how similar cell type composition is between replicates and individuals, we first calculated the proportion of cells from each individual in each replicate assigned to each Seurat cluster at resolution 0.1. Then,
using the ComplexHeatmap R package and performing hierarchical clustering based on the complete linkage method, we visualized the clustering of these replicate-individual groups (Gu, Eils, and Schlesner 2016). We repeated this analysis using Seurat clusters at resolution 0.5, 0.8, and 1 to show that the overall patterns of hierarchical clustering are robust to cluster resolution. We performed an analogous analysis using topic loadings instead of cluster proportions. Here, we determined the loading of each topic on cells from the same individual and replicate, then used the same hierarchical clustering with ComplexHeatmap to visualize patterns of similarity between cells of each individual and replicate (Gu, Eils, and Schlesner 2016).

We also performed hierarchical clustering on gene expression of individual cells. To do so, we took the pseudobulk expression for each individual-replicate-cluster group and filtered to genes expressed in at least 20% of cells in at least one cluster. We then calculated the log10 counts per million expression of each gene. We then generated a heatmap using the ComplexHeatmap R package, again performing hierarchical clustering based on the complete linkage method (Gu, Eils, and Schlesner 2016).

Variance partitioning
Using the same pseudobulk data and precision weights computed by voom from differential expression analysis, we used the VariancePartition R package to quantify the variation attributable to cluster, replicate, and individual (Hoffman and Schadt 2016). We fit a random effect model and modelled cluster, replicate, and individual as random effects. We performed this analysis across all tested Seurat clustering resolutions (0.1, 0.5, 0.8, 1). We performed this analysis using both pseudobulk samples of cells from the same cluster, replicate, and individual and at single cell resolution with each cell as a sample. We also partitioned the variance in each Seurat cluster separately using a random effect model with terms for replicate and individual. For this analysis, we used pseudobulk samples of cells from each individual and replicate.

Power Analysis
To ascertain the power to detect eQTLs and dynamic eQTLs across a range of sample sizes, standardized effect sizes, and experiment sizes. We used a power function as derived in Sarkar et. al. 2019:

$$Pow(\beta, \alpha, n, \sigma) = \Phi \left( \frac{\alpha}{2} \frac{\beta}{\sigma \sqrt{n}} \right)$$

where $\beta$ denotes the true additive significance level, $\alpha$ denotes the significance level, $n$ denotes sample size, and $\sigma$ represents the phenotype standard deviation. This approach assumes a simple linear regression for eQTL mapping and a conservative Bonferroni correction (FWER= 0.05).

To estimate the standard deviation for a given experiment size, we downsampled cells from this experiment, sampling evenly between individuals and replicates to range of experiment sizes from 2,700 cells to 21,600 cells. For each experiment size, we took 10 random samples of cells and calculated pseudobulk expression of cells.
from the same cluster (defined at resolution 1), individual, and replicate. We filtered to include genes expressed in at least 20% of cells in at least one cluster (at resolution 1) in the full set of EB cells. For each sample we then partitioned the median variance attributable to residuals using the `variancePartition` package. We then took the median of the median variance from the 10 samples at each experiment size and fit an exponential decay model to quantify the relationship between experiment size and residual variance. We used square root of this variance to determine the standard deviation for a given experiment size in our power calculations.

**Trajectory inference and identification of dynamic gene modules**

We inferred trajectories using PAGA in Scanpy using Seurat clusters at all tested resolutions (Wolf et al. 2019). We assigned pseudo-time using diffusion pseudo-time with the pluripotent cells assigned as the root (Haghverdi et al. 2016). We then manually traced known developmental trajectories supported by the coarse-grained PAGA graph. At clustering resolution 1, we traced the trajectory from pluripotent cells to hepatocytes (clusters 0, 1, 5, 6, 7, 10, 11, 16, 18, 19, 25, and 22), pluripotent cells to endothelial cells (clusters 0, 1, 4, 5, 6, 7, 11, 16, 18, 22, 24, and 25), and pluripotent cells to neurons (clusters 0, 1, 2, 3, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 15, 26, and 27) (Fig. 6A-C)

We then isolated cells from each of these three trajectories and used Split-GPM to simultaneously cluster samples and identify dynamic gene modules. For this analysis, we divided data into decile pseudo-time bins and calculated pseudobulk gene expression for cells of the same individual, replicate, and pseudo-time bin. We identified 20 dynamic gene modules in each trajectory and interpreted them using gene set enrichment. To understand the variation in dynamic gene expression between individuals and replicates, we re-ran split-GPM ten times and observed how often cells from each individual and replicate were assigned to the same sample cluster.

**Acknowledgements**

We thank all members of the Gilad lab, Sebastian Pott, and Abhishek Sarkar for helpful discussions, and Natalia Gonzales for comments on the manuscript. We thank Sandhiya Arun and the Genomics Core Facility at the University of Chicago for performing 10x collections and sequencing the libraries. Y.G. was funded by NIH grant R35GM131726. K.R. was supported by 5 F31HL146171.

**Competing Interests**

Nothing to disclose

**Data Availability**

All scRNA-seq data have been deposited in the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession number GSE178274.
References


