bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Quantifying the effect of experimental perturbations in single-cell RNA-sequencing data using graph signal processing

Daniel B. Burkhardt[1,†], Jay S. Stanley III[3,†], Ana Luisa Perdigoto[4], Scott A. Gigante[1,2,3], Kevan C. Herold[4], Guy Wolf[5,‡], Antonio J. Giraldez[1,‡], David van Dijk[1,2,‡*], Smita Krishnaswamy[1,2,‡*]

[1]Department of Genetics; [2]Department of Computer Science;
[3]Computational Biology & Bioinformatics Program;
[4]Departments of Immunobiology and Internal Medicine; Yale University, New Haven, CT, USA
[5]Department of Mathematics and Statistics, Université de Montréal, Montreal, QC, Canada

[*]Corresponding authors. E-mail: smita.krishnaswamy@yale.edu, david.vandijk@yale.edu
[†] These authors contributed equally. [‡] These authors contributed equally.

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a powerful tool to quantify transcriptional states in thousands to millions of cells. It is increasingly common for scRNA-seq data to be collected in multiple experimental conditions, yet quantifying differences between scRNA-seq datasets remains an analytical challenge. Previous efforts at quantifying such differences focus on discrete regions of the transcriptional state space such as clusters of cells. Here, we describe a continuous measure of the effect of an experiment across the transcriptomic space. First, we use the manifold assumption to model the cellular state space as a graph (or network) with cells as nodes and edges connecting cells with similar transcriptomic profiles. Next, we create an Enhanced Experimental Signal (EES) that estimates the likelihood of observing cells from each condition at every point in the manifold. We show that the EES has useful properties and information that can be extracted. The EES can be used to identify how gene expression is affected by a given perturbation, including identifying non-monotonic changes from only two conditions. We also show that we can use both the magnitude and frequency of the EES, using an algorithm we call vertex frequency clustering, to derive subsets of cells at appropriate levels of granularity (tailored to areas that change) that are enriched in the experimental or control conditions or that are unaffected between conditions. We demonstrate both algorithms using a combination of biological and synthetic datasets. Implementations are provided in the MELD Python package, which is available at `https://github.com/KrishnaswamyLab/MELD`.

1 Introduction

As single-cell RNA-sequencing (scRNA-seq) has become more accessible, the design of single-cell experiments has become increasingly complex. Researchers regularly use scRNA-seq to quantify the effect of a drug, gene knockout, or other experimental perturbation on a biological system. However, quantifying the compositional differences between single-cell datasets collected from multiple experimental conditions

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remains an analytical challenge [1] because of the heterogeneity and noise in both the data and the effects of a given perturbation.

Previous work has shown the utility of modelling the transcriptomic state space as a continuous low-dimensional manifold, or set of manifolds, to characterize cellular heterogeneity and dynamic biological processes [2–8]. In the manifold model, the biologically valid combinations of gene expression are represented as a smooth, low-dimensional surface in a high dimensional space, such as a two-dimensional sheet embedded in three dimensions. The main challenge in developing tools to quantify compositional differences between single-cell datasets is that each dataset comprises several intrinsic structures of heterogeneous cells, and the effect of the experimental condition could be diffuse or isolated to particular areas of the manifold. Technical noise from scRNA-seq measurements, stochastic biological heterogeneity, and uneven exposure to a perturbation can frustrate any attempts to understand differences between single-cell datasets.

Our goal is to quantify the effect of an experimental perturbation on every single cell state observed in the matched experimental and control scRNA-seq samples of the same biological system. We explicitly define and quantify an *enhanced experimental signal* (EES), which represents the effect of an experimental perturbation across the manifold as a change in the probability of observing each transcriptomic profile in the treatment condition relative to the control. We assume that the cell profiles observed in each experiment are sampled from an underlying multivariate probability density function over the transcriptomic state space that describes which cell states are likely to be observed in a given condition. For example, it is more likely to observe neuronal cells in a sample of brain tissue than in a peripheral blood sample. Next, we assume that the effect of an experimental perturbation is to change this underlying probability density. For example, if you knock out a gene, some neuronal types or even transcriptional states of the same type may be more or less likely to be observed in a scRNA-seq dataset. The key observation here is that we expect to observe a continuous spectrum of changes in probability across the cellular manifold (**Fig. 1**). Because the effect of an experiment is continuous, we seek to estimate this effect across all the observed regions of the manifold, namely at each single-cell profile sampled from either condition.

Although several methods exist for merging multiple single-cell datasets [9, 10], previous work comparing multiple datasets either compare cluster proportions or quantify differential gene expression between samples. Most published analyses of multiple scRNA-seq samples follow the same basic steps [11–18]. First, datasets are merged applying either batch normalization [17, 18] or a simple concatenation of data matrices [11–16]. Next, clusters are identified by grouping either sets of cells or modules of genes. Finally, within each cluster, the cells from each condition are used to calculate statistical measures, such as fold-change between samples. However, reducing the experimental signal to cluster proportions of some fixed size sacrifices the power of single-cell data. In particular, we demonstrate cases in the following sections where subsets of a cluster are enriched and others subsets are depleted, but in the published analysis these nuances were missed because the analysis focused on fold-change in abundance of each cluster.

Instead of quantifying the effect of a perturbation on clusters, we focus on the level of single cells. First, we use the manifold assumption to create a simplified data model, a cell similarity graph where nodes are cells and edges connect cells with similar transcriptomic profiles [19]. We then apply tools from the emerging field of graph signal processing [20] to compute the EES as the likelihood of observing a given cell in the treatment condition relative to the control. This signal takes high values for cell profiles that are more likely to be observed in the experimental condition and less likely to be observed in the control, and vice versa.

In the sections that follow, we show that the EES has useful information for the analysis of experimental conditions in scRNA-seq. First, it can be used as a measure of transcriptional response to a perturbation on

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a cell-by-cell basis to identify the cells most and least affected by an experimental treatment. Second, it is able to identify gene signatures of a perturbation by examining how gene expression covaries with the EES. Third, we show that the frequency composition of the EES can be used as the basis for a clustering algorithm we call *vertex frequency clustering*, which identifies populations of cells that are transcriptionally similar and are similarly affected (either enriched, depleted, or unchanged) between conditions. To demonstrate these advantages, we apply this analysis to a variety of biological datasets, including T-cell receptor stimulation [15], CRISPR mutagenesis in the developing zebrafish embryo [17], and a newly published dataset of interferon-gamma stimulation in human pancreatic islets. We also provide a set of quantitative comparisons for both algorithms using ground truth simulated scRNA-seq data. In each case, we demonstrate the ability of the EES to identify trends across experimental conditions and identify instances where use of the EES and vertex frequency clustering improves over published analytic techniques.

Implementations of the EES algorithm and vertex frequency clustering are provided in the Python package MELD, so named for its utility in joint analysis of single-cell datasets. MELD is open-source and available on GitHub at `https://github.com/KrishnaswamyLab/MELD`.

2 Results

We propose a novel approach to quantifying compositional differences between single-cell experiments inspired by recent successes in applying manifold learning to scRNA-seq analysis [19]. The manifold model is a useful approximation for the cellular transcriptomic space because not all combinations of gene expression are biologically valid. Instead, valid cellular states are intrinsically low-dimensional with smooth transitions between similar states. This implies, for example, that there is no discontinuity in gene expression as a cell transitions between subtypes of the same general cell state within an organism. However, it is possible that the bridges between distant cell states (e.g. between a blood cell and a neuron) are not observed in a dataset because the shared ancestral cell state is transiently present in an earlier developmental stage than that of the experiment. Here, these distant states would be modelled by multiple disconnected manifolds that each are locally continuous. The power of scRNA-seq as a measure of an experimental treatment is that it provides observations of cell state at thousands to millions of points along the manifold in each condition. In this context, our goal is to quantify the change in enrichment of cell states along the manifold as a result of the experimental treatment (**Fig. 1**).

For an intuitive understanding, we first consider a simple experiment with one treatment condition and one control. We seek to compute a score that reflects the conditional likelihood that each cell comes from experimental or control conditions computed over a manifold approximated from all cells from both conditions [19]. This score can be used as a measure of the effect of the experimental treatment because it indicates for each cell how much more likely we are to observe that cell state in the treatment condition relative to the control condition (**Fig. 1**). We refer to this ratio as the *Enhanced Experimental Signal* (EES).

As has been done previously, we approximate the cellular manifold by constructing a simplified data geometry represented by an affinity graph between cells from both conditions [2–8]. In this graph nodes are cells and the edges between nodes describe the transcriptional similarity between the cells. We then take a new approach to analyze the structure of this graph representation inspired by recent advances in graph signal processing [20]. A graph signal is any function that has a defined value for each node in a graph. As such, it is natural to represent gene expression values, labels indicating the sample origin of each each cell, or the EES as a signal over a graph. To derive the EES, first we use the condition from which each cell was sampled to define a signal over the graph that we call the *Raw Experimental Signal* (RES). In a simple two-sample experiment, the RES would be defined as -1 for cells from the control condition and +1 for cells in

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

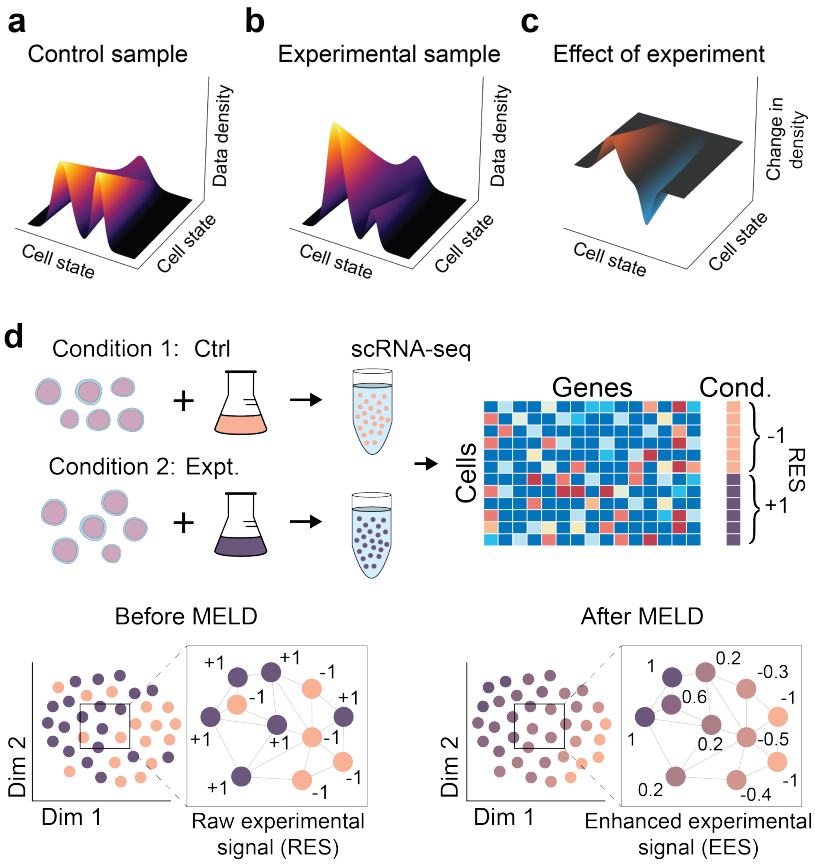

Figure 1: To quantify the effect of an experiment, we consider the results of the control sample (**a**) and experimental samples (**b**) as two empirical probability density estimates over the underlying transcriptomic cell state space. In this context, the experimental effect can be modelled as the change in the probability density in the experiment relative to the control (**c**). (**d**) MELD quantifies this effect by denoising the Raw Experimental Signal (RES) on the cell similarity graph to learn the Enhanced Experimental Signal (EES). The EES indicates how prototypical each cell is of the experimental or control conditions

.

the experimental condition. We then filter the RES by applying a *low pass filter*, which can be thought of as averaging values of the RES across the neighbors on the graph, with higher weighting on nearer neighbors. The output of this filter gives the EES (likelihood score), which is smooth because neighboring cells on the graph will have similar likelihoods of being observed in a given condition. However, because we want the likelihood score to be unaffected by cell sampling variations across the manifold, we use a customized filter that adapts to local variations in density as well as noise patterns in the data. For example, these local variations might represent a small group of cells that are enriched in the experimental condition, but are part of a relatively larger cluster that is depleted.

2.1 Overview of the EES algorithm

We calculate the EES in the following steps that are each explained in more detail below:

1. First, we compute an affinity graph over the cellular state space with an adaptive kernel that cancels out changes in sampling density across the state space.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Next, we create a discrete signal, the RES, over the graph using the labels indicating the sample from which each cell was sequenced.

3. Finally, instead of direct averaging over the cellular state space (or graph domain) to compute likelihood, we apply a novel filter in the frequency domain to smooth this signal.

The first step of the EES algorithm is to create a cell similarity graph in which neighboring cells (i.e., cells with small distances between them) are connected by edges. There are many ways to construct such a graph, and in general the algorithm presented here can work over any such construction. The default graph construction implemented in the MELD toolkit quantifies cell similarity (i.e., the edge weights of the graph) using the α-decay kernel proposed in [3], which can be interpreted as a smooth k-Nearest Neighbors (kNN) kernel. However, in cases where batch normalization is required, we first apply a variant of Mutual Nearest Neighbors (MNN) to merge the datasets [9]. The MNN kernel is described in Section 4.1.11.

Next, we use the input experimental label to create the RES on the graph. For simple two-sample experimental cases, cells from the control condition are assigned a value of -1 and cells from the experimental signal are assigned a value of +1 (**Fig. 2a**). For more complex cases, such as in a series of drug titrations, the raw signal can be assigned continuous values corresponding to the dosage of the drug. Alternatively, the RES can be defined as a multidimensional signal when comparing the differences between categorical conditions, that cannot be defined ordinally on the number line, such as among three or more replicates or genotypes. We discuss this application in Section 2.9.

Finally, to compute likelihood scores we take a weighted average of the RES values at neighborhoods centered at each node on the graph. By taking averages over neighborhoods, we are essentially binning or aggregating information contained within the RES over a region of the graph, similar to how a histogram might be used to estimate density of a variable. This corresponds to a smoothing operation that can be implemented by low-pass filtering in the graph frequency domain. However, as mentioned above, we do not want

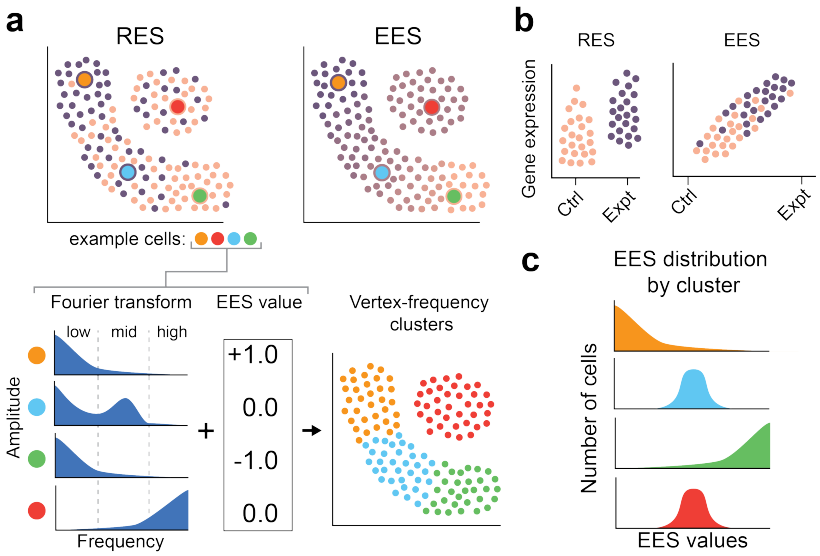

Figure 2: Clustering and Experimental Analysis with MELD (**a**) The Windowed Graph Fourier Transform and EES values at four example points shows distinct patterns between a transitional (blue) and unaffected (red) cell. This information is used for Vertex Frequency Clustering. (**b**) Ordering cells by the EES reveals gene expression changes of the experimental condition. (**c**) Examining the distribution of EES scores in vertex-frequency clusters identifies cell populations most affected by a perturbation.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to apply a simple averaging which might be insensitive to subpopulations of cells being affected differently than neighboring cell states. Instead, we formulate the low-pass filter as an optimization problem where we seek to learn the EES such that it is both smooth signal over the graph and penalized for large changes from the original RES. Thus the resultant low-pass filter respects the changes and decision boundaries offered in the RES while taking local averages at appropriate levels of granularity throughout the manifold to derive an EES that accurately reflects the conditional likelihood of every point in the manifold being generated in the control or experimental conditions.

2.2 Derivation of the EES Low Pass Filter

To derive and explain the custom low-pass filter, we consider the frequency composition of the RES and EES. The analysis of such frequency composition relies on the graph Fourier transform, which is based on a generalization of classic harmonics (i.e., sinusoidal waves oscillating at certain frequencies) to graph harmonics that capture analogous notions or regularity and smoothness [20]. Formally, the graph Fourier transform decomposes graph signals into a weighted sum of eigenvectors of the graph Laplacian, $\mathcal{L}$, which serve as harmonics in this case. For completeness, further details and background from graph signal processing are provided in Section 4.1.3.

Since the EES estimates conditional likelihood of the experimental label over the manifold, we smooth the estimate around a neighborhood of each cell. This operation in the frequency domain corresponds to low-pass filtering. However, we do not use a simple low pass filter for smoothing the EES, which would impose a global standard for smoothness at all points along the manifold. Instead, we allow for variable degrees of smoothness along the manifold by also considering signal reconstruction. Thus, in areas where small populations of cells along of the manifold are enriched or depleted, the filter adapts to these higher frequency changes in the conditional likelihood.

This strategy is expressed as the optimization:

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{y}\|_2^2}_{\text{a}} + \underbrace{\beta \mathbf{y}^T \mathcal{L} \mathbf{y}}_{\text{b}}, \tag{1}$$

Here, $\mathbf{x}$ corresponds to the input RES, $\mathbf{y}^*$ is the desired EES, $\mathbf{y}$ is the set of all possible signals, and $\mathcal{L}$ is a graph Laplacian. The optimization can be broken into two parts: (a) reconstruction, calculated as the minimum Euclidean distance between the raw signal $\mathbf{x}$ and $\mathbf{y}$; and (b) a smoothness penalty that calculates the sum of squared differences of $\mathbf{y}$ across all edges in the graph, adjusted by the weights of the edges. Minimizing these arguments produces the low-pass filtered and denoised signal $\mathbf{y}^*$. Usefully, we find that although this filter is expressed as a convex optimization, it has an exact solution as derived in Section 4.1.7.

Additionally, we want the EES to be robust to technical and experimental noise in the measurements. While the low-pass filter will eliminate "high frequency noise," i.e., adjacent cells on the graph with different values of the RES, it does not address low-frequency noise, which can correspond to a larger biological process that is reflected across the data, such as cell cycle [21]. In these cases, which can be identified by examining gene loadings on Laplacian eigenvectors, we want the flexibility to filter both high-frequency and low-frequency components. To provide this adaptability, we propose a new class of graph filters that can be tuned to graph and signal noise context, given by the following equation:

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{y}\|_2^2 + \mathbf{y}^T \mathcal{L}_* \mathbf{y} \tag{2}$$
$$\text{where } \mathcal{L}_* = [\beta \mathcal{L} - \alpha \mathbf{I}]^\rho .$$

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here, $\mathbf{I}$ is the identity matrix, and α, β, and ρ are parameters that control the spectral translation, reconstruction penalty, and filter order, respectively. In contrast to previous works using Laplacian filters, these parameters allow analysis of signals that are combinations of several underlying changes occurring at various frequencies. For an intuitive example, consider that the frequency of various Google searches will vary depending on if it's winter or summer (low-frequency variation), if it's Saturday or Monday (medium-frequency variation) or it's morning or night (high-frequency variation). In the biological context such changes could manifest as differences in cell type abundance (low-frequency variation) and cell-cycle (medium-frequency variation) [21]. We illustrate such an example in **Fig. S2a** by blindly separating a medium frequency signal from a low frequency contaminating signal over simulated data. Such a technique could be used to separate low- and medium-frequency components so that they can be analyzed independently. We address each of the filter parameters and parameter selection in more detail in Section 4.1.5. For all of the biological datasets and quantitative comparisons presented in this manuscript, we set $\alpha = 0, \beta = 1$, and $\rho = 2$. These are the default parameters implemented in the MELD package.

The optimization in equation (2) may be solved in many ways. To achieve an efficient implementation, the MELD toolkit considers the spectral representation of the RES and uses a Chebyshev polynomial approximation to efficiently compute the EES (see Section 4.1.4). The result is a highly scalable implementation. The EES can be calculated on a dataset of 50,000 cells in less than 8 minutes in a free Google Colaboratory notebook[1], with more than 7 minutes of that spent constructing a graph that can be reused for visualization [3] or imputation [4]. With the EES, it is now possible to address a common problem in single-cell analysis, such as quantifying the effect on an experimental perturbation on gene expression.

2.3 The EES improves inference of differentially expressed genes between conditions

Commonly, one wants to know how gene expression changes between two experimental conditions, i.e., one wants to identify a gene expression signature of a given process. When directly comparing gene expression between samples, the data is organized categorically. This limits analysis to comparison of summary statistics such as mean or variance of gene expression between each category. Furthermore, it is impossible to identify non-linear or non-monotonic changes in gene expression between two samples. One major advantage of applying the EES for analysis of experimental perturbations in scRNA-seq is that the EES is a quantitative vector that varies continuously over the cellular manifold. The cells that are most enriched in each condition have the most extreme EES values, and cells equally likely to be observed in either condition have EES values between the extremes. The continuous nature of the EES makes it possible to order cells by EES values and identify continuous changes in gene expression between the most extreme cell states (**Fig. 2c**).

The EES effectively increases the resolution of the experimental data and enables the recovery of complex non-linear and non-monotonic trends in gene expression with the experimental condition. Even if only two conditions (such as an experiment and control) are measured, the EES can be used to infer which cells exhibit a weak or intermediate response to an experiment. This increased resolution provides the power to regress complex non-linear trends in expression against the EES. We demonstrate the recovery of non-monotonic gene expression signatures on simulated data using only two samples in **Fig. S3**. In this simulated experiment we generate high-dimensional data emulating a biological transition between two terminal cell states through an intermediate transitional population. One of the genes in this simulation has peak expression in the intermediate cell state, but low expression in both terminal states. We show that directly comparing expression of this gene between samples using the RES shows no difference between samples.

[1] Freely available at `colab.research.google.com`, most instances provide a 4-core 2GHz CPU and 20GB of RAM.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, the EES reveals the true pattern of gene expression (**Fig. S3h**).

Beyond examining trends of single genes, one often wants to know which genes are the most strongly affected by an experimental perturbation. These strongly affected genes are often called the gene signature of an experiment or biological process. However, due to technical and biological noise in the experiment, simply calculating fold-change in expression between conditions often fails to recover meaningful changes in gene expression. A key advantage of the EES is that it provides a continuous measure of the experimental signal, which makes it possible to identify gene signatures by ranking genes by their statistical association with the EES (**Fig. 2c**). We previously developed kNN-DREMI (*k*-Nearest Neighbors conditional Density Resampled Estimate of Mutual Information)[22, 23] to quantify such trends in scRNA-seq. To characterize signatures of an experiment, we calculate kNN-DREMI against the EES for all genes and rank the genes by these scores. For example, in Section 2.5, we use this approach to identify the gene signature of T cell activation and show that this signature is enriched for genes known to play a role in activation. It is also possible to quantify changes in expression by calculating fold-change only between the cells that are most enriched in either condition. In Section 2.6, we take this approach to calculate fold-change in expression between cells with the top and bottom 20% of EES values and reveal specific responses within zebrafish cell types to Cas9 mutagenesis of chordin. We anticipate that using the EES to quantify gene signatures of an experiment will be a major use-case for the EES algorithm.

2.4 Vertex-frequency clustering identifies patterns of heterogeneity in high dimensional data

Another common goal for analysis of experimental scRNA-seq data is to identify subpopulations of cells that are responsive to the experimental treatment. Existing methods cluster cells by transcriptome alone and then attempt to quantify the degree to which these clusters are differentially represented in the two conditions. However, this is problematic because the granularity, or sizes, of these clusters may not correspond to the sizes of the cell populations that respond similarly to experimental treatment. Additionally, when partitioning data along a continuum, cluster boundaries are somewhat arbitrary and may not correspond to populations with distinct differences between conditions. Our goal is to identify clusters that are not only transcriptionally similar but also respond similarly to an experimental perturbation.

A naïve approach to identify such clusters would be to simply concatenate the EES to the gene expression data as an additional feature and cluster on these combined features. However, we show that this would not correctly identify subpopulations with respect to their experimental response. **Fig. S4** provides a simulated case for this, generated using a Gaussian mixture model which separates two cell types along Dim 2 based on their responsiveness to a treatment on Dim 1. Traditional analysis may identify two clusters (based on the binary RES); alternatively, clustering based on *k*-means and spectral clustering revealed 4 clusters, and Louvain returned 5 clusters. Each of these clusterings identify the pure populations resulting from the reservoirs of enriched cells in condition 1 and condition 2 (which progress along Dim 1), but each fails to treat the non responsive population appropriately, breaking it into two or three pieces.

However, we conjecture that in this example there are 4 meaningful clusters: two each given by the pure cells from condition 1 and 2, one that is a partition of the purely mixed cells (along Dim 2), and the final cluster is the transitioning population between the two enriched groups of cells. While this is merely an illustrative example, our biological analysis will show that analogous situations occur in real experiments.

As no contemporary method is suitable for finding this transitioning structure, we developed an algorithm that uses the graph Fourier domain to cluster cells based on their latent geometry as well as their EES response (**Fig. S4d**). In particular, we cluster using local frequency profiles of the RES around each cell. This paradigm is motivated by the utility of analyzing cells based on different classes of heterogeneity. This

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

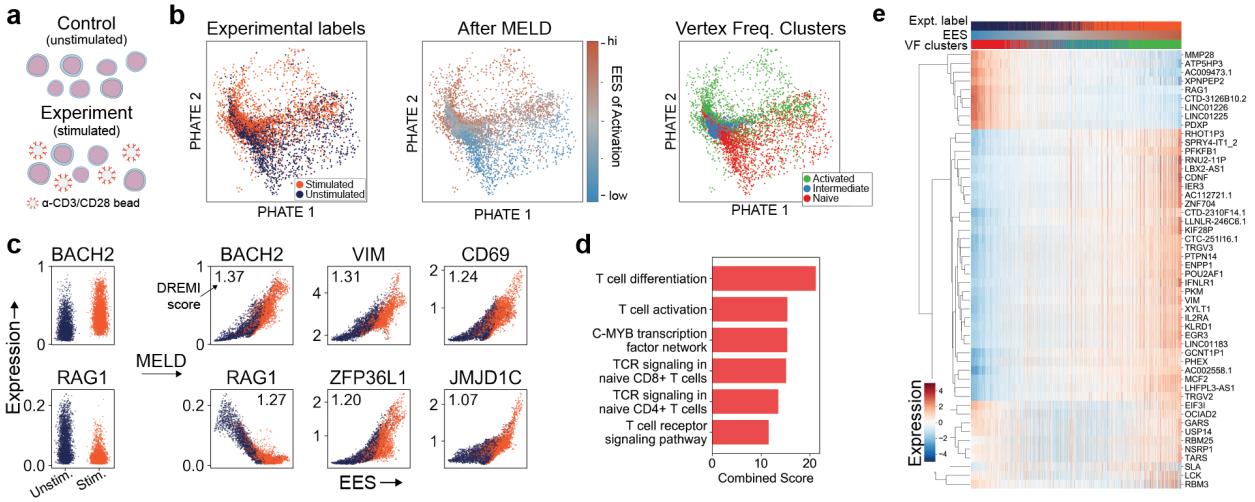

Figure 3: MELD recovers signature of TCR activation. (**a**) Jurkat T-cells were stimulated with α-CD3/CD28 coated beads for 10 days before collection for scRNA-seq. (**b**) Examining a PHATE plot, there is a large degree of overlap in cell state between experimental conditions. However, after MELD it is clear which cells states are prototypical of each experimental condition. (**c**) Relationship between gene expression and TCR activation state is revealed when cells are ordered by the EES instead of grouped by experimental condition. (**d**) Signature genes identified by top 1% of kNN-DREMI scores are enriched for annotations related to TCR activation. (**e**) Z-scored expression of select signature genes ordered by the EES reveals patterns of up- and downregulation. Notice a subset of genes exhibit non-monotonic expression patterns, such as USP14 and NSRP1. Identifying such trends would be impossible without MELD.

method, which we call *vertex-frequency clustering*, is an adaptation of the signal-biased spectral clustering proposed by Shuman et al. [24].

Briefly, the method considers sums of many scales of spectrograms generated from the RES. Each spectrogram is obtained by translating a window function that considers neighborhoods of a specific scale at each vertex, then taking the resulting Fourier transform of the windowed signal. The result of this operation is a vertex-frequency matrix that is N-cells by N-frequencies. Each scale is then activated using a nonlinear transformation and summed. Finally, the summation is concatenated with the EES vector, and k-means is used to cluster the cells based on their multiscale vertex-frequency characteristics. Vertex-frequency clustering separates the value in the EES from its spectral characteristics and allows one to consider both the local spectra as well as the signal value. By considering both vertex and frequency information, one may distinguish between heterogeneous populations which are non-responsive and heterogeneous populations which are in transition.

The algorithm briefly proposed above is discussed in further detail in methods Section 4.2. In particular, we detail a fast implementation using the recently proposed fast graph Fourier transform [25] and the diffusion operator. In the following sections, we demonstrate EES filtering and vertex-frequency clustering on biological data.

2.5 The EES identifies a biologically relevant signature of T cell activation

To demonstrate the ability of the EES to identify a biologically relevant EES, we apply the algorithm to 5,740 Jurkat T cells cultured for 10 days with and without anti-CD3/anti-CD28 antibodies published by Datlinger et al. [15] (**Fig. 3a**). The goal of this experiment was to characterize the transcriptional signature of T cell Receptor (TCR) activation. We visualize the data using PHATE, a visualization and dimensionality

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reduction tool we developed for single-cell RNA-seq data (**Fig. 3b**)[3]. We observe a large degree of overlap in cell states between the stimulated and control conditions, as noted in the original study[15]. This overlap has both technical and biological causes. Approximately 76% of the cells were transfected with gRNAs targeting proteins in the TCR pathway, meaning that the remaining 24% of cells in the stimulated condition lack key effectors of the activation pathway and are expected to resemble naive cells. The expectation is for these cells to appear transcriptionally naive despite originating from the stimulated experimental condition. In other words, although the RES for these cells is +1 (originating from the stimulated condition), the EES of these cells is expected to be closer to -1 (prototypical of the unstimulated condition).

To obtain a signature of T cell activation, Datlinger et al. [15] devised an *ad hoc* iterative clustering approach whereby cells were first clustered by the gRNA observed in that cell and then further clustered by the gene targeted. In each cluster, the median gene expression was calculated and the first principle component was used as the dimension of activation. The 165 genes with the highest component loadings were defined as signature genes and used to judge the level of activation in each cell. In contrast, using the EES analysis we can derive a gene signature of TCR activation at single-cell resolution without relying on clustering or access to information about the gRNA observed in each cell.

Applying the EES algorithm to the data, we observe a continuous spectrum of scores across the dataset (**Fig. 3b**). As expected, the regions enriched for cells from the stimulated condition have higher EES values representing highly activated cells, and the converse is true for regions enriched for unstimulated cells. To ensure that the EES represents a biologically relevant dimension of activation, we generate a gene signature comparable to the results of Datlinger et al. [15] by selecting genes with a high mutual information with the EES using kNN-DREMI[4, 22]. We then perform gene set enrichment analysis on the top 165 genes using EnrichR[26] (**Fig. 3c,e**). We find comparable enrichment for gene sets related to T cell activation, T cell differentiation, and TCR response (**Fig. 3d**) and identify an overlap of 53 genes between the EES-inferred and published signatures. We find that in the GO sets of T cell activation, T cell differentiation, and T cell receptor signalling, the EES signatures includes as many or more genes for each GO term. Furthermore, our signature includes genes known to be affected by TCR stimulation but not present in the Datlinger et al. [15] signature list, such as down regulation of RAG1 and RAG2 [27]. These results show that the EES is capable of identifying a biologically relevant dimension of T cell activation at the resolution of single cells without relying on knowledge of the treatment through the gRNAs as in Datlinger et al. [15].

2.6 Characterizing genetic loss-of-function mutations in the developing zebrafish

To demonstrate the utility of EES analysis applied to complex datasets composed of multiple cell types, we applied EES analysis to a recently published chordin loss-of-function experiment in zebrafish using CRISPR/Cas9 (**Fig. 4**)[17]. In this system, loss of chordin function results in a ventralization phenotype characterized by expansion of the ventral mesodermal tissues at the expense of the dorsally-derived neural tissues[28–30]. In Wagner et al. [17], zebrafish embryos were injected at the 1-cell stage with Cas9 and gRNAs targeting either chordin (*chd*), a BMP-antagonist required for developmental patterning, or tyrosinase (*tyr*), a control gene required for pigmentation but not expected to affect cell composition at these stages. Embryos were collected for scRNA-seq at 14-16 hours post-fertilization (hpf). Similar to the T cell dataset above, we expect incomplete penetrance of the perturbation in this dataset because not all cells in the experimental condition will successfully receive the gRNA needed to cause the loss-of-function mutation.

To characterize the effect of chordin mutagenesis, Wagner et al. [17] projected cells from each sample onto 28 clusters obtained from a reference wild-type dataset. Within each cluster, the fold-change of cells from the *tyr*-injected to *chd*-injected condition was calculated and MAST[12] was used to calculate differentially expressed genes. A drawback of this approach is the restriction of analysis of the experimental

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

effect to clusters, instead of single cells. This means that there is no way to detect divergent responses across subpopulations within clusters. Here, we demonstrate the ability of the EES to detect such occurrences and show how VF clustering detects groups of cells with similar responses to an experimental perturbation.

First, we derived an EES of response to chordin loss-of-function. Here, cells with high EES values correspond to cells prototypical of the *chd* samples and low EES values correspond to cells prototypical of the *tyr* samples (**Fig. 4a**). To identify the effect of mutagenesis on various cell populations, we first examined the distribution of EES scores across the 28 cell state clusters generated by Wagner et al. [17] for this dataset (**Fig. 4b**). As expected, we find that Mesoderm – Lateral Plate (MLP), Tailbud – Presomitic Mesoderm (TPM), Hatching Gland (HG), and Mesoderm – Blood Island (MBI) have the highest average EES values, matching the observed expansion of the mesoderm and blood tissues in the embryos injected with *chd* gRNAs [17]. The cells with the lowest EES values are the Optic Primordium (OP), Differentiating Neurons (DN), Neural – Diencephalon (NDI), and Notochord (NTC). This is interpreted as finding these tissues in a *tyr* embryo, but not in a *chd* embryo, matching observed deficiencies of these tissues in the absence of chordin[28–30]. These results confirm that the EES is able to identify the effect of experimental perturbations across many cell types.

2.7 VF clustering identifies subpopulations in the Tailbud - Presomitic Mesoderm cluster

An advantage of using the EES instead of fold-change is the ability to examine the distribution of scores within a cluster to understand the heterogeneity of the response. In analyzing the chordin loss-of-function experiment, we observe that the Tailbud – Presomitic Mesoderm (TPM) cluster exhibits the largest range of EES values. This large range suggests that there are cells in this cluster with many different responses to *chd* mutatgenesis. To investigate this effect further, we generated a PHATE plot of the cluster (**Fig. 4c**). In this visualization, we observe many different branches of cell states, each with varying ranges of EES values. We use vertex-frequency clustering to identify clusters of cells that are transcriptionally similar and exhibit a homogeneous response to perturbation (**Fig. 4d**).

Within the PSM cluster, we find four subclusters. Using established markers[18], we identify these clusters as immature adaxial cells, mature adaxial cells, the presomitic mesoderm, and forming somites (**Fig. 4c, S5**). Examining the distribution of EES scores within each cell type, we conclude that the large range of EES values within the TPM cluster is due to largely non-overlapping distributions of scores within each of these subpopulations (**Fig. 4e**). The mature and immature adaxial cells, which are muscle precursors, have low EES values. This indicates depletion of these cells in the *chd* condition which matches observed depletion of myotomal cells in chordin mutants[28]. Conversely, the presomitic mesoderm and forming somites have high EES values, indicating that these cells are prototypically enriched in a chordin mutant. Indeed, expansion of these presomitic tissues is observed in siblings of the *chd* embryos[17]. This heterogeneous effect was entirely missed by the fold-change analysis, since the averaging of all cells assigned to the PSM cluster masked the depletion of adaxial cells.

Another advantage of vertex-frequency clustering is that we can now calculate differential expression of genes within these populations of cells that we infer have homogeneous responses to a perturbation. Examining the distribution of genes within each of the identified subclusters, we find different trends in expression within each group (**Fig. 4f**). For example, Myod1, a marker of adaxial cells, is lowly expressed

[2]Abbreviations: MLP: Lateral plate, TPM: Tailbud - Presomitic mesoderm, HG: Hatching gland, MBI: Blood island, EPP: Epidermal - pfn1, MEN: Endothelial, PRD: Periderm, EPA: Epidermal anterior, EPO: Otic placode, LLP: Lateral line, EPF: Epidermal - foxi3a, GL: Germline, NRB: Rohon beard, NFP: Floorplate, MHF: Heart field, MPA: Pharyngeal arch, NCC: Neural crest - crestin, END: Endoderm, TSC: Tailbud - spinal cord, NC: Neural crest, NTE: Telencephalon, MPD: Pronephric duct, NHB: Hindbrain, NMB: Midbrain, NTC: Notocord, NDI: Diencephalon, DN: Neurons, OP: Optic

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

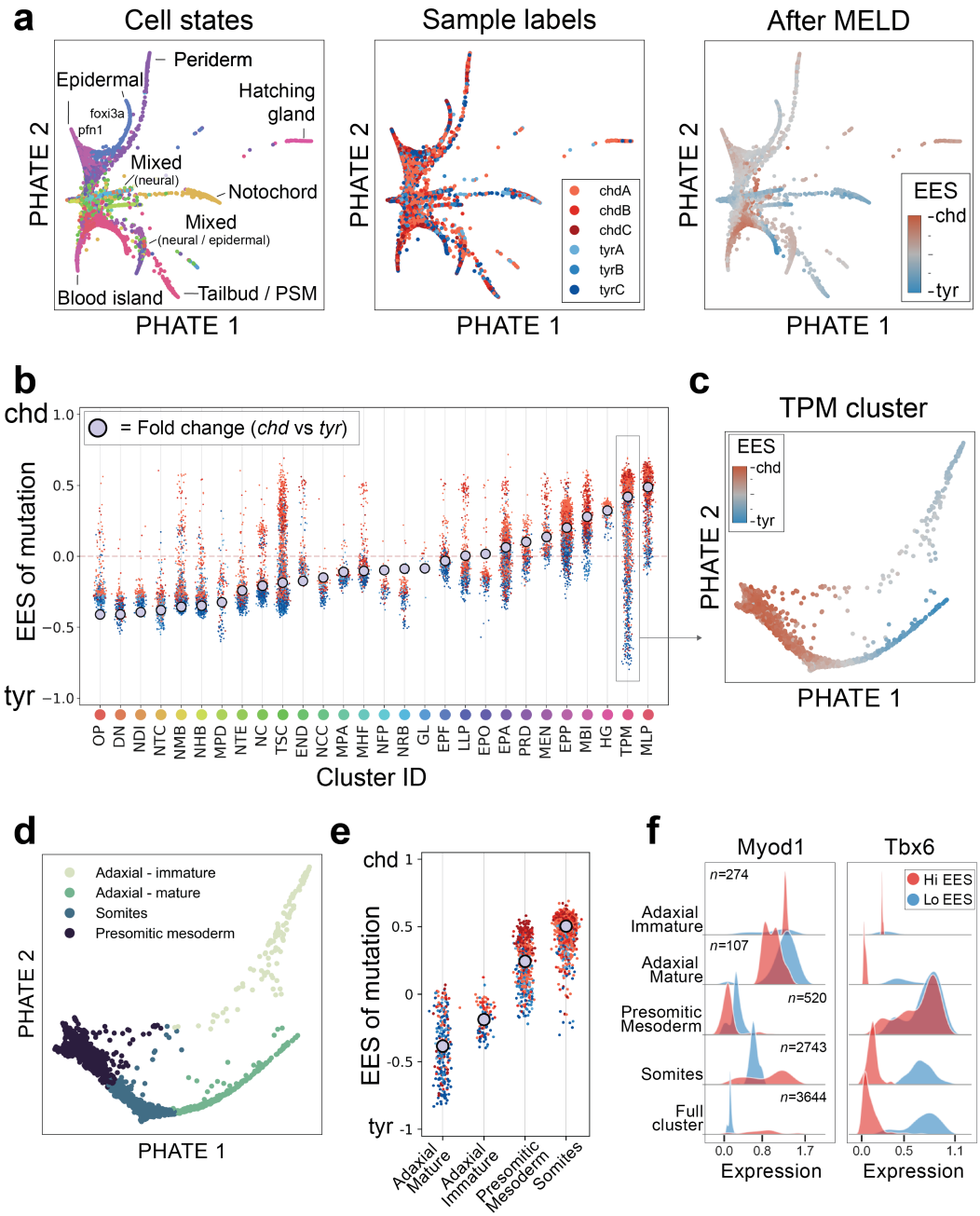

Figure 4: Characterizing chordin Cas9 mutagenesis with MELD. (**a**) PHATE shows a high degree of overlap of sample labels across cell types. Applying MELD to the mutagenesis vector reveals regions of cell states enriched in the *chd* or *tyr* conditions. (**b**) Using published cluster assignments[2], we show that the EES quantifies the effect of the experimental perturbation on each cell, providing more information than calculating fold-change in the number of cells between conditions in each cluster (grey dot), as was done in the published analysis. Color of each point corresponds to the sample labels in panel (a). Generally, average EES value aligns with the fold-change metric. However, we can identify clusters, such as the TPM or TSC, with large ranges of EES values indicating non-uniform response to the perturbation. (**c**) Visualizing the TPM cluster using PHATE, we observe several cell states with mostly non-overlapping EES values. (**d**) Vertex Frequency Clustering identifies four cell types in the TPM. (**e**) We see the range of EES values in the TPM cluster is due to subpopulations with divergent responses to the *chd* perturbation. (**f**) Changes in gene expression within subclusters is lost when only considering the full cluster, as was done in the published analysis.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the presomitic mesoderm and in the somites, but highly expressed in the adaxial cells. Attempting to compare the difference in expression of this gene in the entire cluster would be obfuscated by differences in abundance of each cell subpopulation between samples. We find a similar trend with Tbx6, a marker of the presomitic mesoderm, which is not expressed in adaxial cells and mature somites (**Fig. 4f**). Using the EES analysis, we observe that Tbx6 expression of presomitic mesoderm cells is unchanged in the *chd* mutants whereas analysis of the cluster considered by Wagner et al. [17] would suggest a strong change in expression. With EES and vertex frequency clustering analysis, we can see that the observed change in the published analysis was due to Simpson's paradox, where changes in abundance of some subpopulations lead to misleading differences in statistics calculated across multiple populations as a whole. Note also that if we had merely compared the fold-change in abundance in the *chd* vs *tyr* conditions, as was done in the published analysis, we would have completely missed this effect and instead only observed that there is a 2-fold change in abundance of this cluster between samples. These results demonstrate the advantage of using the EES and vertex frequency clustering to quantify the effect of genetic loss-of-function perturbations in a complex system with many cell types.

2.8 Identifying the effect of IFNγ stimulation on pancreatic islet cells

Next we use the EES to characterize a newly generated dataset of human pancreatic islet cells cultured for 24 hours with and without interferon-gamma (IFNγ), a system with significant clinical relevance to auto-immune diseases of the pancreas such as Type I Diabetes mellitus (T1D). The pathogenesis of T1D is generally understood to be caused by T cell mediated destruction of beta cells in the pancreatic islets[31] and previous reports suggest that islet-infiltrating T cells secrete IFNγ during the onset of T1D[32]. It has also been described that IFNγ-expressing T cells mediate rejection of pancreatic islet allografts[33]. Previous studies have characterized the effect of these cytokines on pancreatic beta cells using bulk RNA-sequencing[34], but no studies have addressed this system at single-cell resolution.

To better understand the effect of immune cytokines on islet cells, we cultured islet cells from three donors for 24 hours with and without IFNγ and collected cells for scRNA-seq. After filtering, we obtain 5,708 cells for further analysis. Examining the expression of marker genes for major cell types of the pancreas, we observe a noticeable batch effect associated with the donor ID, driven by the maximum expression of glucagon, insulin, and somatostatin in alpha, beta, and delta cells respectively (**Fig. S6a**). To correct for this difference while preserving the relevant differences between samples, we apply the MNN kernel correction described in Section 4.1.11 to merge cells from each donor. Examining PHATE plots after batch correction, we observe three distinct populations of cells corresponding to alpha, beta, and delta cells (**Fig. 5a**).

To quantify the effect of IFNγ treatment across these cell types, we calculate the EES of IFNγ stimulation(**Fig. 5a**). We then apply vertex-frequency clustering to identify nine subpopulations of cells. Using established marker genes of islet cells[35], we determine that these clusters correspond to alpha, beta, and delta cells (**Fig. 5a,b**, **Fig. S6b**). We first characterize the gene expression signature of IFNγ treatment across these cell types. Using kNN-DREMI[4] to identify genes with a strong association with the EES, we observe strong activation of genes in the JAK-STAT pathway including STAT1 and IRF1[36] and in the IFN-mediated antiviral response including MX1, OAS3, ISG20, and RSAD2 [37–39] (**Fig. 5c**). The activation of both of these pathways has been previously reported in beta cells in response to IFNγ[40, 41]. Furthermore, we observe a high degree of overlap in the IFNγ response between alpha and beta cells, but less so between delta cells and either alpha or beta cells. Examining the genes with the top 1% of kNN-DREMI scores (n=196), we find 62 shared genes in the signatures of alpha and beta cells, but only 22 shared by alpha, beta, and delta cells. To confirm the validity of our gene signatures, we use EnrichR[26] to perform

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

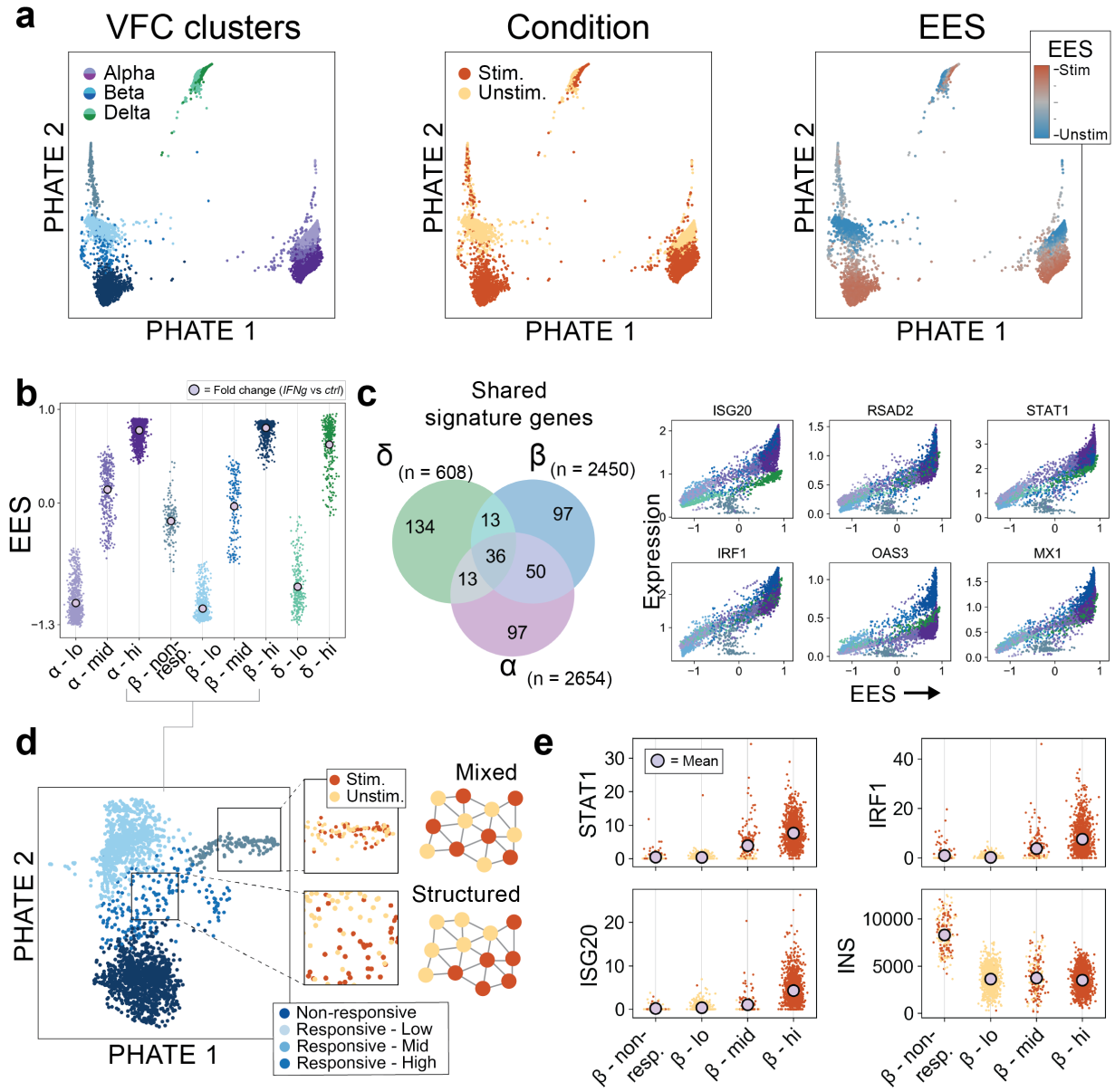

Figure 5: MELD characterizes the response to IFNγ in pancreatic islet cells. (**a**) PHATE visualization of pancreatic islet cells cultured for 24 hours with or without IFNγ. Vertex-frequency clustering identifies nine clusters corresponding to alpha, beta, and delta cells. (**b**) Examining the EES in each cluster, we observe that beta cells have a wider range of responses than alpha or delta cells. (**c**) We identify the signature of IFNγ stimulation by calculating kNN-DREMI scores of each gene with the EES. We find a high degree of overlap of the top 1% of genes by kNN-DREMI score between alpha and beta cells. (**d**) Examining the four beta cell clusters more closely, we observe two populations with intermediate EES values. These populations are differentiated by the structure of the RES in each cluster (outset). In the non-responsive cluster, the RES has very high frequency unlike the low frequency pattern in the transitional Responsive - mid cluster. (**e**) We find that the non-responsive cluster has low expression of IFNγ-regulated genes such as STAT1 despite containing roughly equal numbers of unstimulated (n=123) and stimulated cells (n=146). This cluster is marked by approximately 2.5-fold higher expression of insulin.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene set enrichment analysis on the 196 signature genes and find strong enrichment for terms associated with interferon signalling pathways (**Fig. S6c**). From these results we conclude that although IFNγ leads to upregulation of the canonical signalling pathways in all three cell types, the response to stimulation in delta cells is subtly different to that of alpha or beta cells.

We next examine the distribution of EES values within each of the clusters identified by vertex-frequency clustering (**Fig. 5b**). Interestingly, choosing $k = 9$ clusters, we find two clusters of beta cells with intermediate EES values. These clusters are cleanly separated on the PHATE plot of all islet cells (**Fig. 5a**) and together represent the largest range of EES scores in the dataset. To further inspect these clusters, we consider a separate PHATE plot of the cells in the four beta cell clusters (**Fig. 5d**). Examining the distribution of RES values in these intermediate cell types, we find that one cluster, which we label as *Non-responsive*, exhibits high frequency RES values indicative of a population of cells that does not respond to an experimental treatment (**Fig. 5d** - outset). The *Responsive - Mid* cluster matches our characterization of a transitional population with a structured distribution of RES values. Supporting this characterization, we find a lack of upregulation in IFNγ-regulated genes such as STAT1 in the non-responsive cluster, similar to the cluster of beta cells with the lowest EES values (**Fig. 5e**).

In order to understand the difference between the non-responsive beta cells and the responsive populations, we calculate differential expression of genes in the non-responsive clusters and all others as previously described [4]. The gene with the greatest difference in expression is insulin, the marker of beta cells, which is approximately 2.5-fold increased in the non-responsive cells (**Fig. 5e**). This cluster of cells bears resemblance to a recently described "extreme" population of beta cells that exhibit elevated insulin mRNA levels and are found to be more abundant in diabetic mice[42, 43]. That these cells appear non-waresponsive to IFNγ stimulation and exhibit extreme expression of insulin suggests that the presence of extreme high insulin in a beta cell prior to IFNγ exposure may inhibit the IFNγ response pathway through an unknown mechanism.

Here, we applied EES analysis to a new dataset to identify the signature of IFNγ stimulation across alpha, beta, and delta cells. Furthermore, we used vertex frequency clustering to identify a population of beta cells with high insulin expression that appears unaffected by IFNγ stimulation. Together, these results demonstrate the utility of EES analysis to reveal novel biological insights in a clinically-relevant biological experiment.

2.9 Analysis of donor-specific composition

Although most of the analysis in this manuscript focuses on the two-sample condition, we show that it is possible to use the EES to quantify the differences between more than two conditions. In the islet dataset, we have samples of treatment and control scRNA-seq data from three different donors. To quantify the differences in cell profiles between samples, we first use a one-hot matrix to create three RES vectors. That is, the first vector has value 1 associated with cells that were sampled from donor 1 islets and is 0 elsewhere, the second vector has value 1 associated with cells that were sampled from donor 2 islets and 0 elsewhere, and so on. We then use the EES algorithm to smooth each RES independently. This produces a measure of how likely each cell's transcriptional profile is to be observed in donor 1, 2, or 3. We then analyze each of these signals for each cluster examined in Section 2.8 (**Supp. Fig. S7**). We find that all of the alpha cell and delta cell clusters are depleted in donor 3 and the non-responsive beta cell cluster is enriched primarily in donor 1. Furthermore, the most highly activated alpha cell cluster is enriched in donor 2. As with the EES derived for the IFNγ response, it is also possible to identify donor-specific changes in gene expression, or clusters of cells differentially abundant between each donor.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.10 Quantitative comparisons using simulated data

To demonstrate the accuracy of the EES algorithm and vertex frequency clustering, we designed a set of quantitative comparisons using simulated scRNA-seq data. To generate datasets with known ground-truth, we use Splatter, a package for simulating scRNA-seq data with a specified geometry [44]. We designed four base dataset structures using a mixture of branching trajectories and discrete clusters. For each of these datasets, we created a ground-truth likelihood ratio defining the probability that each cell would be assigned to one of two conditions. The scRNA-seq expression values and sign, magnitude, and size of the regions of enrichment and depletion are randomly generated during each simulation. **Fig. 6a** shows one representative simulation from each of the four dataset structures.

To quantify the accuracy of the EES algorithm in recovering the ground-truth experimental signal, we compare the EES algorithm to two approaches to smoothing signals defined over graphs. The first is k-nearest neighbor averaging, which has been used in denoising images [45] and gene expression values [46]. We also consider averaging the RES over the neighbors of the graph, which is one of the simplest low-pass filters for graph signals [47]. For these experiments, we generated a total of 120 Splatter datasets using the 4 base geometries. We created the ground truth probability that each cell would be observed in the experimental or control condition and generated sample labels and RES for each cell according to these probabilities. Finally, we ran each algorithm on the RES and calculated the Pearson Correlation between the output signal and the likelihood ratios used to generate each dataset. On average, the EES algorithm outperforms both kNN averaging and the graph averaging by 17% and 32%, respectively (**Fig. 6b**).

Next, we quantify the ability of vertex frequency clustering to identify the regions of each dataset that are enriched, depleted, or unchanged in the experimental condition relative to the control. Here, we compared vertex frequency clustering to KMeans and Spectral Clustering (as implemented in Scikit-learn [48]) and Phenograph [6]. We ran the simulations as described above and calculated the Adjusted Rand Score between the clusters identified from each method and the ground truth data partitions used to generate the ground truth experimental signal. Although vertex frequency clustering performs best on average across all methods, there is a much larger variation in performance between runs of the same dataset in these quantification compared to the EES comparisons. Examining the non-monotonic branch dataset, we find that this variation is related to changes in the relative sizes of the enriched regions. When the differentially abundant regions of the data were relatively even proportions, all algorithms perform similarly, and when the differentially abundant regions are especially large or small, then all algorithms perform poorly (**Fig. S8**). Nonetheless, vertex frequency clustering outperforms all three competing methods for all considered values.

Overall, these comparisons demonstrate the utility of EES and vertex frequency clustering analysis across a variety of pseudo-biological data geometries. We note that the accuracy of these methods, especially vertex frequency clustering, vary as the size of the differentially abundant region changes. Additionally, it is important to note that all methods were run using default parameters for these comparisons. We encourage future comparison of the EES algorithm and vertex frequency clustering to future methods for quantifying compositional differences in scRNA-seq datasets as new tools are developed.

3 Discussion

When performing multiple scRNA-seq experiments in various experimental and control conditions, researchers often seek to characterize the cell types or sets of genes that change from one condition to another. However, quantifying these differences is challenging due to the subtlety of most biological effects relative to the biological and technical noise inherent to single-cell data. To overcome this hurdle, we designed the

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

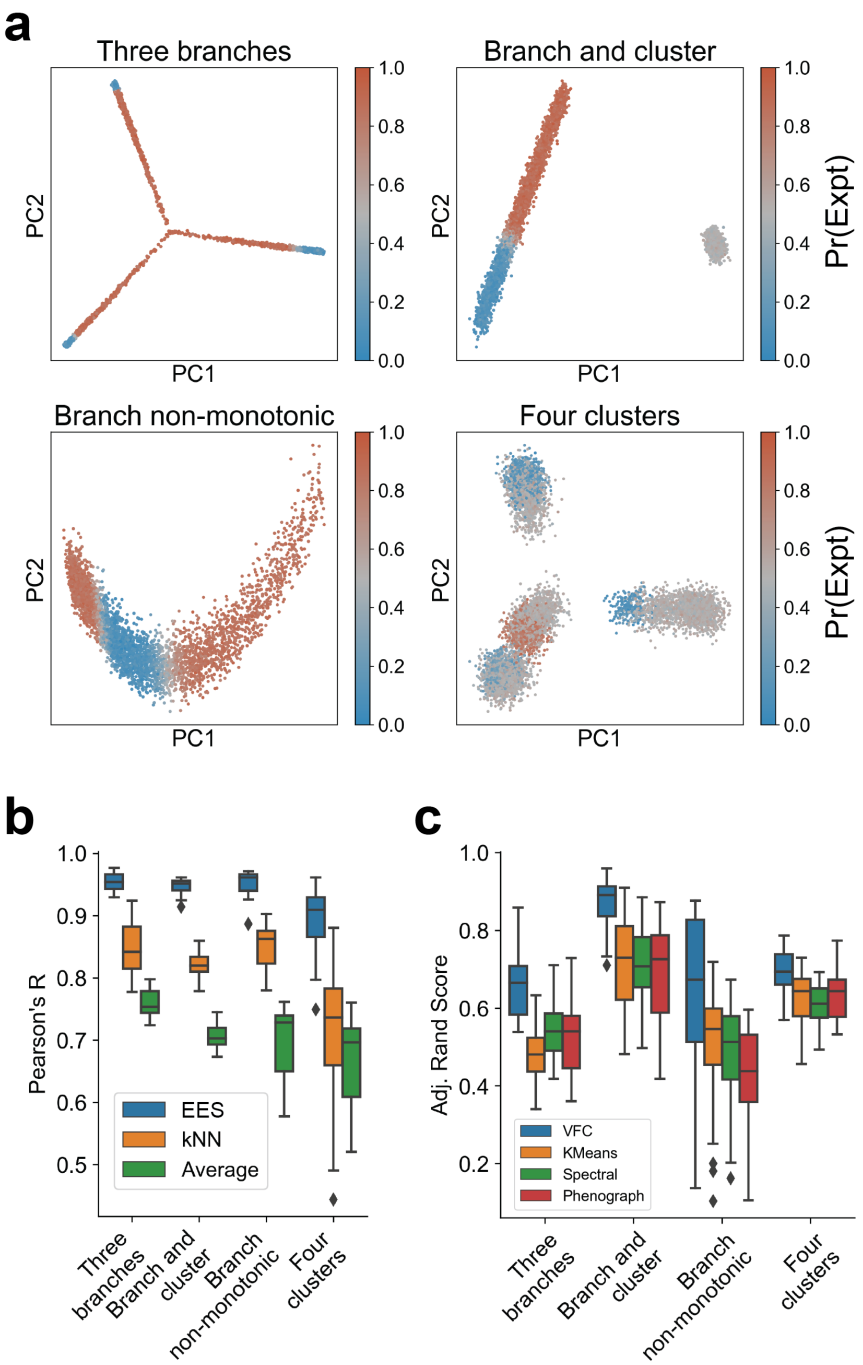

Figure 6: Quantitative comparison of the EES and VFC. (**a**) Single cell datasets were generated using Splatter [44]. Each cell is colored by the probability that cell would be observed in the experimental condition relative to the control. The data and ground truth probabilities were randomly generated 20 times with varying noise and regions of enrichment. (**b**) Comparison of the EES algorithm to kNN averaging of the RES and graph Averaging. (**c**) Comparison of VFC to popular clustering algorithms. Adjusted Rand Score quantifies how accurately each method detects regions that were enriched, depleted, or unchanged in the experimental condition relative to the control.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EES algorithm and vertex frequency clustering to quantify compositional differences between samples.

The EES can be used to identify individual cells that are the most likely to be observed in each sample and can be used to identify changes in gene expression between conditions. The EES can also be used to identify groups of cells that do not change between experimental conditions. We demonstrate that using the EES, it is possible to identify non-linear and non-monotonic changes in gene expression that would be lost through a direct comparison of expression between two samples. These benefits can be applied to experimental designs of two or more categorical condition labels.

We show in Section 2.6 that EES analysis improves over the current best-practice strategy of clustering cells based on gene expression and calculating differential abundance and differential expression within clusters. Clustering prior to quantifying compositional differences can fail to identify the divergent responses of subpopulations of cells within a cluster. To identify clusters of cells with cohesive responses to a perturbation, we introduce a novel clustering algorithm, called Vertex-Frequency Clustering. Using the RES and EES, we derive clusters of cells as the correct cluster size to identify cells that are most enriched in either condition, cells transitioning between these states, and cells that are unaffected by an experimental perturbation. The applications of EES and vertex frequency clustering analysis are demonstrated on single-cell datasets from three different biological systems and experimental designs. We also provide quantitative comparisons of the EES algorithm and vertex frequency clustering using simulated scRNA-seq data with known ground truth. To facilitate the application of these tools for future scRNA-seq analysis, we provide open-source Python implementations that inherit the Scikit-learn API in the MELD package on GitHub `https://github.com/KrishnaswamyLab/MELD`.

The flexibility of EES analysis and vertex frequency clustering to analyze arbitrary signals over a cell similarity graph suggest several future applications in scRNA-seq analysis. For example, in **Fig. S2** we demonstrate the ability of analysis with the MELD toolkit to extract convoluted signals of different frequencies on a graph. These two signals might represent a cell cycle effect, experimental signal, and technical noise. By tracking genes that vary with cell cycle, for example, we could remove this trend from the experiment to improve the identification of gene signatures of an experimental perturbation. Another potential application of MELD is the comparison of multiple experimental meta-variables. One can imagine an experiment where cells are exposed to combinations of drugs in varying concentrations with the goal of understanding how these combinations of drugs interact. By building a unified cell similarity graph across conditions, one could deconvolve the signals of each component of the treatment and then calculate a measure of association, such as mutual information, to identify which drugs elicit similar or divergent effects alone or in combination. This flexibility makes MELD an ideal analytical tool for scRNA-seq experiments across biological systems.

4 Computational Methods

In this section, we will provide details about our computational methods for computing the EES, as well as extracting information from the EES by way of a method we call *vertex frequency clustering*. We will outline the mathematical foundations for each algorithm, explain how they relate to previous works in manifold learning and graph signal processing, and provide details of the implementations of each algorithm.

4.1 Computation of the EES

Computing the EES involves the following steps each of which we will describe in detail.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. A cell similarity graph is built over the combined data from all samples where each node or vertex in the graph is a cell and edges in the graph connect cells with similar gene expression values.

2. The condition label for each cell is used to create the Raw Experimental Signal (RES).

3. The RES is then smoothed over the graph to calculate the EES using a graph filter called the EES filter.

4.1.1 Graph construction

The first step in the EES algorithm is to create a cell similarity graph. In single-cell RNA sequencing, each cell is measured as a vector of gene expression counts measured as unique molecules of mRNA. Following best practices for scRNA-seq analysis [1], we normalize these counts by the total number of Unique Molecular Indicators (UMIs) per cell to give relative abundance of each gene and apply a square-root transform. Next we compute the similarity all pairs of cells, by using their Euclidean distances as an input to a kernel function. More formally, we compute a similarity matrix W such that each entry W_{ij} encodes the similarity between cell gene expression vectors $\mathbf{x_i}$ and $\mathbf{x_j}$ from the dataset X.

In our implementation we use α-decaying kernel proposed by Moon et al. [3] because in practice it provides an effective graph construction for scRNA-seq analysis. However, in cases where batch, density, and technical artifacts confound graph construction, we also use a mutual nearest neighbor kernel as proposed by Haghverdi et al. [9].

The α-decaying kernel [3] is defined as

$$W_{i,j}=\frac{1}{2}\exp\left(-\left(\frac{\|x_i-x_j\|_2}{\varepsilon_k(x_i)}\right)^{\alpha}\right)+\frac{1}{2}\exp\left(-\left(\frac{\|x_i-y_i\|_2}{\varepsilon_k(x_j)}\right)^{\alpha}\right), \tag{3}$$

where x_i, y_i are data points, $\varepsilon_k(x_i), \varepsilon_k(x_j)$ are the distance from x_i, x_j to their k-th nearest neighbors, respectively, and α is a parameter that controls the decay rate (i.e., heaviness of the tails) of the kernel. This construction generalizes the popular Gaussian kernel, which is typically used in manifold learning, but also has some disadvantages alleviated by the α-decaying kernel, as explained in Moon et al. [3].

The similarity matrix effectively defines a weighted and fully connected graph between cells such that every two cells are connected and that the connection between cells x_i and x_j is given by $W(i, j)$. To allow for computational efficiency, we sparsify the graph by setting very small edge weights to 0.

4.1.2 Estimating density versus conditional likelihood on a graph

There has been a body of literature that addresses performing high dimensional density estimation scalably using a graph representation of the data [49–53]. Instead of estimating kernel density or histograms in N dimensions where N could be large, these methods rendered the data as a graph, and density is estimated each point on the graph (each data point) as some variant counting the number of points which lie within a radius $r - steps$ of each point on the graph.

While we compare to methods that perform local aggregation of information in Section 2.10, we use a more complex formulation for estimating a *conditional likelihood of the experimental label* rather than a density. To make this distinction clear, we do not compute density, *i.e.* estimate of how many points from each condition are in each neighborhood on the manifold. This would be confounded by differences in cell number and sampling density between the experiments. Instead, conditioned on being in each location on the graph, we compute how likely it is that the given cell was generated in the conditional or experimental condition. Thus we aim to eliminate the effect of absolute density differences along the manifold and focus

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on changes in local likelihood between the two conditions. This is achieved by way of the α-decay kernel described above, which effectively adapts the radius r above to the density of the neighborhood.

An approach that would eliminate density and compute likelihood would be to use an adaptive bandwidth kernel and compute the ratio of cells with RES value +1 or -1 in these adapted neighborhoods. However, instead we formulate a convex optimization that balances smoothing (or local aggregation) of the RES into the EES based on the adaptive bandwidth, and respecting trends in the original RES. This is because we want to infer the conditional likelihood as smoothly varying (so as to denoise likelihood signal), but do not want to require the estimate have the same smoothness globally across the graph. Instead, we want to allow for local variations in smoothness. This allows the EES to capture changes in density at multiple scales including small areas of the manifold that may be either enriched or depleted between conditions.

Finally, we want the inferred conditional likelihood to be robust to noise. Thus, we allow for bandwidth-adjustable filtering of frequencies in the RES to be eliminated as noise. If the user believes the signal has low frequency noise the optimization can be used to eliminate low frequencies as well as high frequencies.

In the next few sections, we will explain how we set up the computation of the EES as a convex optimization which derives a signal over the the cell-similarity graph using tools of graph signal processing.

4.1.3 Graph Signal Processing

The EES algorithm leverages recent advances in graph signal processing (GSP) [20], which aim to extend traditional signal processing tools from the spatiotemporal domain to the graph domain. Such extensions includes, for example, wavelet transforms[54], windowed Fourier transforms [24], and uncertainty principles [55]. All of these extensions rely heavily on the fundamental analogy between classical Fourier transform and graph Fourier transform (described in the next section) derived from eigenfunctions of the graph Laplacian, which is defined as

$$\mathcal{L} := D - W, \tag{4}$$

where D is the *degree* matrix, which is a diagonal matrix with $D_{ii} = d(i) = \sum_j^N W_{ij}$ containing the degrees of the vertices of the graph defined by W.

4.1.4 The Graph Fourier Transform

One of the fundamental tools in traditional signal processing is the Fourier transform, which extracts the frequency content of spatiotemporal signals[56]. Frequency information enables various insights into important characteristics of analyzed signals, such as pitch in audio signals or edges and textures in images. Common to all of these is the relation between frequency and notions of *smoothness*. Intuitively, a function is *smooth* if one is unlikely to encounter a dramatic change in value across neighboring points. A simple way to imagine this is to look at the *zero-crossings* of a function. Consider, for example, sine waves $\sin ax$ of various frequencies $a = 2^k, k \in \mathbb{N}$. For $k = 0$, the wave crosses the x-axis (a zero-crossing) when $x = \pi$. When we double the frequency at $k = 1$, our wave is now twice as likely to cross the zero and is thus less smooth than $k = 0$. This simple zero-crossing intuition for smoothness is relatively powerful, as we will see shortly.

Next, we show that our notions of smoothness and frequency are readily applicable to data that is not regularly structured, such as single-cell data. The graph Laplacian $\mathcal{L}$ can be considered as a graph analog of the Laplace (second derivative) operator ∇^2 from multivariate calculus. This relation can be verified by deriving the graph Laplacian from first principles.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For a graph $\mathcal{G}$ on N vertices, its graph Laplacian $\mathcal{L}$ and an arbitrary graph signal $\mathbf{f} \in \mathbb{R}^N$, we use equation (4) to write

$$\begin{aligned}(\mathcal{L}\,\mathbf{f})\,(i) &= ([D - W]\,\mathbf{f})\,(i) \\ &= d(i)\mathbf{f}(i) - \sum_j W_{ij}\mathbf{f}(j) \\ &= \sum_j W_{ij}\,(\mathbf{f}(i) - \mathbf{f}(j))\,. \end{aligned} \tag{5}$$

As the graph Laplacian is a weighted sum of differences of a function around a vertex, we may interpret it analogously to its continuous counterpart as the curvature of a graph signal. Another common interpretation made explicit by derivation (5) is that $(\mathcal{L}\mathbf{f})(i)$ measures the *local variation* of a function at vertex i.

Local variation naturally leads to the notion of *total variation*,

$$\mathbf{TV}(\mathbf{f}) = \sum_{i,j} W_{ij}(\mathbf{f}(i) - \mathbf{f}(j))^2,$$

which is effectively a sum of all local variations. $\mathbf{TV}(\mathbf{f})$ describes the global smoothness of the graph signal $\mathbf{f}$. In this setting, the more smooth a function is, the lower the value of the variation. This quantity is more fundamentally known as the *Laplacian quadratic form*,

$$\mathbf{f}^T\mathcal{L}\,\mathbf{f} = \sum_{i,j} W_{ij}(\mathbf{f}(i) - \mathbf{f}(j))^2. \tag{6}$$

Thus, the graph Laplacian can be used as an operator and in a quadratic form to measure the smoothness of a function defined over a graph. One effective tool for analyzing such operators is to examine their eigensystems. In our case, we consider the eigendecomposition $\mathcal{L} = \Psi\Lambda\Psi^{-1}$, with eigenvalues[3] $\Lambda := \{0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N\}$ and corresponding eigenvectors $\Psi := \{\psi_i\}_{i=1}^N$. As the Laplacian is a square, symmetric matrix, the spectral theorem tells us that its eigenvectors in Ψ form an orthonormal basis for $\mathbb{R}^N$. Furthermore, the Courant-Fischer theorem establishes that the eigenvalues in Λ are local minima of $\mathbf{f}^T\mathcal{L}\mathbf{f}$ when $\mathbf{f}^T\mathbf{f} = 1$ and $\mathbf{f} \in U$ as $\dim(U) = i = 1, 2, \cdots N$. At each eigenvalue λ_i this function has $\mathbf{f} = \psi_i$. In summary, the eigenvectors of the graph Laplacian (1) are an orthonormal basis and (2) minimize the Laplacian quadratic form for a given dimension.

Henceforth, we use the term *graph Fourier basis* interchangeably with graph Laplacian eigenvectors, as this basis can be thought of as an extension of the classical Fourier modes to irregular domains[20]. In particular, the ring graph eigenbasis is composed of sinusoidal eigenvectors, as they converge to discrete Fourier modes in one dimension. The graph Fourier basis thus allows one to define the *graph Fourier transform* (GFT) by direct analogy to the classical Fourier transform.

The GFT of a signal f is given by $\hat{f}(\lambda_\ell) = \sum_i f(i)\psi_\ell^T(i) = \langle \mathbf{f}, \psi_\ell \rangle$, which can also be written as the matrix-vector product

$$\hat{\mathbf{f}} = \Psi^T\mathbf{f}. \tag{7}$$

As this transformation is unitary, the inverse graph Fourier transform (IGFT) is $\mathbf{f} = \Psi\hat{\mathbf{f}}$. Although the graph setting presents a new set of challenges for signal processing, many classical signal processing notions

[3]Note that in this discussion we abuse notation by treating Λ as an ordered set of Laplacian eigenvalues and as the diagonal matrix with entries from the elements of this set. Similarly, Ψ is both the set of column eigenvectors $\{\psi_i\}_{i=1}^N$ as well as the $N \times N$ matrix $[\psi_1\psi_2 \cdots \psi_N]$ with eigenvector as a column.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such as filterbanks and wavelets have been extended to graphs using the GFT. We use the GFT to process, analyze, and cluster experimental signals from single-cell data using a novel graph filter construction and a new harmonic clustering method.

4.1.5 The EES Filter

In the EES algorithm, we seek to estimate the change in likelihood between two experimental labels along a manifold represented by a cell similarity graph. To estimate likelihood along the graph, we employ a novel graph filter construction, which we explain in the following sections. To begin, we review the notion of filtering with focus on graphs, and demonstrate the filter in a low-pass setting. Next, we demonstrate the expanded version of the EES filter and provide an analysis of its parameters. Finally, we provide a simple solution to the EES filter that allows fast computation.

4.1.6 Filters on graphs

In their simplest forms, filters can be thought of as devices that alter the spectrum of their input. Filters can be used as bases, as is the case with wavelets, and they can be used to directly manipulate signals by changing the frequency response of the filter. For example, many audio devices contain an equalizer that allows one to change the amplitude of bass and treble frequencies. Simple equalizers can be built simply by using a set of filters called a filterbank. In the EES algorithm, we use a tunable filter to amplify latent features on a single-cell graph.

Mathematically, graph filters work analogously to classical filters. Particularly, a filter takes in a signal and attenuates it according to a frequency response function. This function accepts frequencies and returns a response coefficient. This is then multiplied by the input Fourier coefficient at the corresponding frequency. The entire filter operation is thus a reweighting of the input Fourier coefficients. In low-pass filters, the function only preserves frequency components below a threshold. Conversely, high-pass filters work by removing frequencies below a threshold. Bandpass filters transfer frequency components that are within a certain range of a central frequency. The tunable filter in the EES algorithm is capable of producing any of these responses.

As graph harmonics are defined on the set Λ, it is common to define them as functions of the form $h : [0, \max(\Lambda)] \mapsto [0, 1]$. For example, a low pass filter with cutoff at λ_k would have $h(x) > 0$ for $x < \lambda_k$ and $h(x) = 0$ otherwise. By abuse of notation, we will refer to the diagonal matrix with the filter h applied to each Laplacian eigenvalue as $h(\Lambda)$, though h is not a set-valued or matrix-valued function. Filtering a signal $\mathbf{f}$ is clearest in the spectral domain, where one simply takes the multiplication $\hat{\mathbf{f}}_{\text{filt}} = h(\Lambda)\hat{\mathbf{f}} = h(\Lambda)\Psi^*\mathbf{f}$.

Finally, it is worth using the above definitions to define a vertex-valued operator to perform filtering. As a graph filter is merely a reweighting of the graph Fourier basis, one can construct the *filter matrix*,

$$H = \Psi h(\Lambda)\Psi^T. \tag{8}$$

A simple manipulation using equation (7) will verify that $H\mathbf{f}$ is the WGFT of $\hat{\mathbf{f}}_{\text{filt}}$. This filter matrix will be used to solve the EES filter in approximate form for computational efficiency.

4.1.7 Laplacian Regularization

A simple assumption for recovering a the conditional likelihood EES signal from raw measurements is *smoothness*. In this model the latent signal is assumed to have a low amount of neighbor to neighbor

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variation. *Laplacian regularization*[57–65] is a simple technique that targets signal smoothness via the optimization

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\text{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\text{b}}. \tag{9}$$

Note that this optimization has two terms, the second is the term that ensures smoothness of the signal. However, note that the first term is called a *reconstruction penalty*, aims to keep the EES similar to the RES. This term will help adjust the amount of smoothness achieved, by the amount of overall smoothness available in the RES.

Laplacian regularization is a sub-problem of the EES filter that we will discuss for low-pass filtering. In the above, a reconstruction penalty (a) is considered alongside the Laplacian quadratic form (b), which is weighted by the parameter β. The Laplacian quadratic form may also be considered as the norm of the *graph gradient*, i.e.

$$\beta \mathbf{z}^T \mathcal{L} \mathbf{z} = \beta \|\nabla_G \mathbf{z}\|_2^2.$$

Thus one may view Laplacian regularization as a minimization of the edge-derivatives of a function while preserving a reconstruction. Because of this form, this technique has been cast as *Tikhonov regularization*[59, 66], which is a common regularization to enforce a high-pass filter to solve inverse problems in regression. In our results we demonstrate a EES filter that may be reduced to Laplacian regularization using a squared Laplacian.

In section 4.1.6 we introduced filters as functions defined over the Laplacian eigenvalues ($h(\Lambda)$) or as vertex operators (equation 8). Minimizing optimization 9 reveals a similar form for Laplacian regularization. Although the of the EES filter is presented as an optimization, we find that it has an exact solution. To begin,

$$\begin{aligned}
\mathbf{y} &= \underset{\mathbf{z}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{z}\|_2^2 + \beta \mathbf{z}^T \mathcal{L} \mathbf{z} \\
&= \underset{\mathbf{z}}{\operatorname{argmin}} (\mathbf{x} - \mathbf{z})^T (\mathbf{x} - \mathbf{z}) + \beta \mathbf{z}^T \mathcal{L} \mathbf{z} \\
&= \underset{\mathbf{z}}{\operatorname{argmin}} \mathbf{x}^T \mathbf{x} + \mathbf{z}^T \mathbf{z} - 2\mathbf{x}^T \mathbf{z} + \beta \mathbf{z}^T \mathcal{L} \mathbf{z}
\end{aligned}$$

Substituting $y = z$, we next differentiate with respect to y and set this to 0,

$$\begin{aligned}
0 &= \nabla_{\mathbf{y}} (\mathbf{x}^T \mathbf{x} + \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{x} + \beta \mathbf{y}^T \mathcal{L} \mathbf{y}) \\
&= 2\mathbf{y} - 2\mathbf{x} + 2\beta \mathcal{L} \mathbf{y} \\
\mathbf{x} &= (\mathbf{I} + \beta \mathcal{L}) \mathbf{y},
\end{aligned}$$

so the solution to problem 9 is

$$\mathbf{y} = (\mathbf{I} + \beta \mathcal{L})^{-1} \mathbf{x}. \tag{10}$$

As the input x is a graph signal in the vertex domain, the least squares solution (10) is a filter matrix $H_{\text{reg}} = (I + \beta \mathcal{L})^{-1}$ as discussed in section 4.1.6. The spectral properties of Laplacian regularization immediately follow as

$$\begin{aligned}
H_{\text{reg}} &= (\mathbf{I} + \beta \mathcal{L})^{-1} \\
&= \Psi \frac{1}{1 + \beta \Lambda} \Psi^T.
\end{aligned} \tag{11}$$

Thus Laplacian regularization is a graph filter with frequency response $h_{\text{reg}}(\lambda) = (1 + \beta\lambda)^{-1}$. Figure S2b shows that this function is a low-pass filter on the Laplacian eigenvalues with cutoff parameterized by β.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.1.8 Tunable Filtering

Though simple low-pass filtering with Laplacian regularization is a powerful tool for many machine learning tasks, we sought to develop a filter that is flexible and capable of filtering the signal at any frequency. To accomplish these goals, we introduce the EES filter:

$$\mathbf{y} = \underset{z}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mathbf{z}^T \mathcal{L}_* \mathbf{z} \qquad (12)$$
$$\text{where } \mathcal{L}_* = [\beta\mathcal{L} - \alpha\mathbf{I}]^\rho .$$

This filter expands upon Laplacian regularization by the addition of a new smoothness structure. Early and related work proposed the use of a power Laplacian smoothness matrix S in a similar manner as we apply here[59], but little work has since proven its utility. In our construction, α is referred to as modulation, β acts as a reconstruction penalty, and ρ is filter order. These parameters add a great deal of versatility to the EES filter, and we demonstrate their spectral and vertex effects in Figure **S2**, as well as provide mathematical analysis of the EES algorithm parameters in section 4.1.9. Finally, in section 4.1.10 we discuss an implementation of the filter.

4.1.9 Parameter Analysis

A similar derivation as section 4.1.7 reveals the filter matrix

$$H_{\text{EES}}(\lambda) = [\mathbf{I} + (\beta\mathcal{L} - \alpha\mathbf{I})^\rho]^{-1} . \qquad (13)$$

which has the frequency response

$$h_{\text{EES}}(\lambda) = \frac{1}{1 + (\beta\lambda - \alpha)^\rho}. \qquad (14)$$

Thus, the value of the EES algorithm parameters in the vertex optimization (12) has a direct effect on the graph Fourier domain. First, we note by inspection that $h_{\text{EES}}(\lambda) = h_{\text{reg}}(\lambda)$ for $\alpha = 0$ and $\rho = 1$ (see equation 11). Thus the EES filter is a superset of graph filters in which Laplacian regularization is a special case.

It is clear that β acts analogously in (14) as it does in the subfilter (11). In each setting, β steepens the cutoff of the filter and shifts it more towards its central frequency (**Fig. S2b**). In the case of $\alpha = 0$, this frequency is $\lambda_1 = 0$. This is done by scaling all frequencies by a factor of β. For stability reasons, we choose $\beta > 0$, as a negative choice of β yields a high frequency amplifier.

The parameters α and ρ change the filter from low pass to band pass or high pass. Figure **S2** highlights the effect on frequency response of the filters and showcases their vertex effects in simple examples. We begin our mathematical analysis with the effects of ρ.

ρ powers the Laplacian harmonics. This steepens the frequency response around the central frequency of the EES filter and, for even values, makes the function square-integrable. Higher values of ρ lead to sharper tails (**Fig. S2c, S2e**), limiting the frequency response outside of the target band, but with increased response within the band. For technical reasons we do not consider odd-valued $\rho > 1$ when $\alpha > 0$ or $\rho \notin \mathbb{N}$. Indeed, though the parameters β and α do not disrupt the definiteness of $\mathcal{L}_*$ (thus $\mathcal{L}_*$ is defined for $\rho \notin \mathbb{N}$), odd-valued and fractional matrix powers of $\mathcal{L}_*$ result in hyperbolic and unstable filter discontinuities. When $\alpha = 0$, these discontinuities are present only at $\lambda = 0$ and are thus stable. However, when $\alpha > 0$, the hyperbolic behavior of the filter is unstable as these discontinuities now lie within the Laplacian spectrum. Finally, ρ can be used to make a high pass filter by setting it to negative values (**Fig. S2f**).

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the integer powers used in EES, a basic vertex interpretation of ρ is available. Each column of $\mathcal{L}^k$ is *k-hop localized*, meaning that $\mathcal{L}^k_{ij}$ is non-zero if and only if the there exists a path length k between vertex i and vertex j (for a detailed discussion of this property, see Hammond et al. [54, section 5.2].) Thus, for $\rho \in \mathbb{N}$, the operator $\mathcal{L}^\rho$ considers variation over a hop distance of ρ. This naturally leads to the spectral behavior we demonstrate in Figure **S2c**, as signals are required to be smooth over longer hop distances when $\alpha = 0$ and $\rho > 1$.

The parameter α removes values from the diagonal of $\mathcal{L}$. This results in a *modulation* of frequency response by translating the Laplacian harmonic that yields the minimal value for problem (12). This allows one to change the target frequency when $\rho > 1$, as α effectively modulates a band-pass filter. As graph frequencies are positive, we do not consider $\alpha < 0$. In the vertex domain, the effect of α is more nuanced. We study this parameter for $\alpha > 0$ by considering a modified Laplacian $\mathcal{L}_*$ with $\rho = 1$. However, due to hyperbolic spectral behavior for odd-valued ρ, $\alpha > 0$ is ill-performing in practice, so this analysis is merely for intuitive purposes, as similar results extend for $\rho > 1$.

For mathematical analysis of α, $\mathcal{L}_*$ is applied as an operator (equation 5) to an arbitrary graph signal $\mathbf{f}$ defined on a graph G. Expanding $(\mathcal{L}_*\mathbf{f})(i)$ we have the following

$$\begin{aligned}
(\mathcal{L}_*\mathbf{f})(i) &= ([\beta(D-W) - \alpha\mathbf{I}]\,\mathbf{f})(i) \\
&= \beta(D\mathbf{f} - W\mathbf{f} - \frac{\alpha}{\beta}\mathbf{f})(i) \\
&= \beta\left[(d(i) - \frac{\alpha}{\beta})\mathbf{f}(i) - \sum_j W_{ij}\mathbf{f}(j)\right] \\
&= \beta\left[\sum_j (W_{ij} - \frac{\alpha}{N\beta})\mathbf{f}(i) - \sum_j W_{ij}\mathbf{f}(j)\right] \\
&= \beta\sum_j W_{ij}\left[(1 - \frac{\alpha}{d(i)\beta})\mathbf{f}(i) - \mathbf{f}(j)\right]. \qquad (15)
\end{aligned}$$

Relation (15) establishes the vertex domain effect of α, which corresponds to a reweighting of the local variation at vertex i by a factor of $1 - \frac{\alpha}{d(i)\beta}$. The intuition that follows is that positive α allows disparate values of $\mathbf{f}$ around each vertex to minimize problem (12), which leads to greater response for high frequency harmonics. We demonstrate this modulation in Figure **S2d**.

To conclude, we propose a filter parameterized by reconstruction β (**Fig. S2b**), order ρ (**Fig. S2c, S2e**), and modulation α (**Fig. S2d**). The parameters α and β are limited to be strictly greater than or equal to 0. When $\alpha = 0$, ρ may be any integer, and it adds more low-frequencies to the frequency response as it becomes more positive. On the other hand, if ρ is negative and $\alpha = 0$, ρ controls a high pass filter. When $\alpha > 0$, ρ must be even-valued and the EES filter becomes a band-pass filter. In standard use cases we propose to use the parameters $\alpha = 0, \beta = 1,$ and $\rho = 2$. All of our biological results were obtained using this parameter set, which gives a square-integrable low-pass filter. As these parameters have direct spectral effects, their implementation in an efficient graph filter is straightforward and presented in section 4.1.10.

4.1.10 Implementation

A naive implementation of the EES algorithm would apply the matrix inversion presented in equation 13. This approach is untenable for the large single-cell graphs that the EES algorithm is designed for, as H^{-1}_{EES} will have many elements, and, for high powers of ρ or non-sparse graphs, extremely dense. A second

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

approach to solving Equation 12 would diagonalize $\mathcal{L}$ such that the filter function in Equation 14 could be applied directly to the Fourier transform of input raw experimental signals. This approach has similar shortcomings as eigendecomposition is substantively similar to inversion. Finally, a speedier approach might be to use conjugate gradient or proximal methods. In practice, we found that these methods are not well-suited for EES filtering.

Instead of gradient methods, we use Chebyshev polynomial approximations of $h_{\text{EES}}(\lambda)$ to rapidly approximate and apply the EES filter. These approximations, proposed by Hammond et al. [54] and Shuman et al. [67], have gained traction in the graph signal processing community for their efficiency and simplicity. Briefly, a truncated and shifted Chebyshev polynomial approximation is fit to the frequency response of a graph filter. For analysis, the approximating polynomials are applied as polynomials of the Laplacian multiplied by the signal to be filtered. As Chebvyshev polynomials are given by a recurrence relation, the approximation procedure reduces to a computationally efficient series of matrix-vector multiplications. For a more detailed treatment one may refer to Hammond et al. [54] where the polynomials are proposed for graph filters. For application of the EES filter to a small set of input RES, Chebyshev approximations offer the simplest and most efficient implementation of our proposed algorithm. For sufficiently large sets of RES, such as when considering hundreds of conditions, the computational cost of obtaining the Fourier basis directly may be less than repeated application of the approximation operator; in these cases, we diagonalize the Laplacian either approximately through randomized SVD or exactly using eigendecomposition, depending on user preference. Then, one simply constructs $H_{\text{EES}} = \Psi h_{\text{EES}}(\Lambda)\Psi^T$ to calculate the EES from the RES.

4.1.11 Addressing batch effects using a Mutual Nearest Neighbor kernel

While the kernel in Eqn. 3 provides an effective way of capturing neighborhood structure in data, it is susceptible to batch effects. For example, when data is collected from multiple patients, subjects, or environments (generally referred to as "batches"), such batch effects can cause affinities within each batch are often much higher than between batches, thus artificially creating separation between them rather than follow the underlying biological state. To alleviate such effects, we adjust the kernel construction using an approach inspired by recent work from by Haghverdi et al. [9] on the Mutual Nearest Neighbors (MNN) kernel. We extend the standard MNN approach, which has previous been applied to the k-Nearest Neighbors kernel, to the α-decay kernel as follows. First, within each batch, the affinities are computed using (3). Then, across batches, we compute slightly modified affinities as

$$K'_{k,\alpha}(x,y)=\min\left\{\exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon'_k(x)}\right)^{\alpha}\right),\exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon'_k(y)}\right)^{\alpha}\right)\right\},$$

where $\varepsilon'_k(x)$ are now computed via the k-th nearest neighbor of x in the batch containing y (and vice versa for $\varepsilon'_k(y)$). Next, a rescaling factor γ_{xy} is computed such that

$$\sum_{z\in\text{batch}(y)} \gamma_{xy} K'_{k,\alpha}(x,z) \leq \beta \sum_{z\in\text{batch}(x)} K_{k,\alpha}(x,z)$$

for every x and y, where $\beta > 0$ is a user configurable parameter. This factor gives rise to the rescaled kernel

$$K'_{k,\alpha,\beta}(x,y) = \begin{cases} K'_{k,\alpha}(x,y) & \text{if batch}(x) = \text{batch}(y) \\ \gamma_{xy} K'_{k,\alpha}(x,y) & \text{otherwise.} \end{cases}$$

Finally, the full kernel is then computed as

$$K'_{k,\alpha}(x,y) = \min\left\{K'_{k,\alpha,\beta}(x,y), K'_{k,\alpha,\beta}(y,x)\right\},$$

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and used to set the weight matrix for the constructed graph over the data. Notice that this construction is a well defined extension of (3), as it reduces back to that kernel when only a single batch exists in the data.

4.1.12 Summary of the EES algorithm

In summary, we have proposed a family of graph filters based on a generalization of Laplacian regularization framework to implement the computation of the EES. This optimization, which we are able to solve analytically allows us to derive the EES, or conditional likelihood of the experimental label, as a smooth and denoised signal, while also respecting multi-resolution changes in the likelihood landscape. As we show in Section 2.10, this formulation performs better at deriving the true conditional likelihood in simulated scRNA-seq data with known ground truth. Further, it is efficient to compute.

The EES algorithm is implemented in Python 3 as part of the MELD package and is built atop the `scprep`, `graphtools`, and `pygsp` packages. We developed `scprep` efficiently process single-cell data, and `graphtools` was developed for construction and manipulation of graphs built on data. Fourier analysis and Chebyshev approximations are implemented using functions from the `pygsp` toolbox[68]. These packages are available through the `pip` package manager. MELD is available on GitHub at `https://github.com/Kri` and on `pip` as `meld`.

4.2 Vertex-frequency clustering

Next, we will describe the vertex frequency clustering algorithm for partitioning the cellular manifold into regions of similar response to experimental perturbation. For this purpose, we use a technique proposed in Shuman et al. [24] based on a graph generalization of the classical Short Time Fourier Transform (STFT). This generalization will allow us to simultaneously localize signals in both frequency and vertex domains. The output of this transform will be a spectrogram Q, where the value in each entry $Q_{i,j}$ indicates the degree to which the RES in the neighborhood around vertex i is composed of frequency j. We then concatenate the EES and perform k-means clustering. The resultant clusters will have similar transcriptomic profiles, similar EES values, and similar *frequency trends* of the RES. The frequency trends of the RES are important because they allow us to infer movements in the cellular state space that occur during experimental perturbation.

We derive vertex frequency clusters in the following steps:

1. We create the cell graph in the same way as is done to derive the EES in Section 4.1.1.
2. For each vertex in the graph (corresponding to a cell in the data), we create a series of localized windowed signals by masking the RES using a series of heat kernels centered at the vertex. Graph Fourier decomposition of these localized windows capture frequency of the RES at different scales around each vertex.
3. The graph Fourier representation of the localized windowed signals is thresholded using a $tanh$ activation function to produce pseudo-binary signals.
4. These pseudo-binarized signals are summed across windows of various scales to produce a single $N \times N$ spectrogram Q. PCA is performed on the spectrogram for dimensionality reduction.
5. The EES is concatenated to the reduced spectrogram weighted by the $L2$-norm of PC1 to produce $\hat{Q}$ which captures both local RES frequency trends and changes in conditional density around each cell in both datasets.
6. k-Means is performed on the concatenated matrix to produce vertex-frequency clusters.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.2.1 Analyzing frequency content of the RES

Before we go into further detail about the algorithm, it may be useful to provide some intuitive explanations for why the frequency content of the RES provides a useful basis for identifying clusters of cells affected by an experimental perturbation. Because the low frequency eigenvectors of the graph Laplacian identify smoothly varying axes of variance through a graph, we associate trends in the RES associated these low-frequency eigenvectors as biological transitions between cell states. This may correspond to the shift in T cells from naive to activated, for example. We note that at intermediate cell transcriptomic states between the extreme states that are most enriched in either condition, we observe both low and middle frequency RES components, see the blue cell in the cartoon in **Fig 2a**. This is because locally, the RES varies from cell to cell, but on a large scale is varying from enriched in one condition to being enriched in the other. This is distinct from what we observe in our model when a group of cells are completely unaffected by an experimental perturbation. Here, we expect to find only high frequency variations in the RES and no underlying transition or low-frequency component. The goal of vertex frequency clustering is to distinguish between these four cases: enriched in the experiment, enriched in the control, intermediate transitional states, and unaffected populations of cells. We also want these clusters to have variable size so that even small groups of cells that may be differentially abundant are captured in our clusters.

4.2.2 Using the Windowed Graph Fourier Transform (WGFT) to identify local changes in RES frequency

While the graph Fourier transform is useful for exploring the frequency content of a signal, it is unable to identify how the frequency content of graph signals change locally over different regions of the graph. In vertex frequency clustering, we are interested in understanding how the frequency content of the RES changes in neighborhoods around each cell. In the time domain, the windowed Fourier transform identifies changing frequency composition of a signal over time by taking slices of the signal (e.g. a sliding window of 10 seconds) and applying a Fourier decomposition to each window independently (WFT) [56]. The result is a spectrogram Q, where the value in each cell $Q_{i,j}$ indicates the degree to which time-slice i is composed of frequency j. Recent works in GSP have generalized the constructions windowed Fourier transform to graph signals[24]. To extend the notion of a sliding window to the graph domain, Shuman et al. [24] write the operation of translation in terms of convolution as follows.

The *generalized translation operator* $T_i : \mathbb{R}^N \to \mathbb{R}^N$ of signal f to vertex $i \in \{1, 2, ..., N\}$ is given by

$$(T_i f)(n) := \sqrt{N}(f * \delta_i)(n), \quad \delta_i(j) = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases} \tag{16}$$

which convolves the signal f, in our case the RES, with a dirac at vertex i. Shuman et al. [24] demonstrate that this operator inherits various properties of its classical counterpart; however, the operator is not isometric and is affected by the graph that it is built on. Furthermore, for signals that are not tightly localized in the vertex domain and on graphs that are not directly related to Fourier harmonics (e.g., the circle graph), it is not clear what graph translation implies.

In addition to translation, a *generalized modulation operator* is defined by Shuman et al. [24] as $M_k : \mathbb{R}^N \to \mathbb{R}^N$ for frequencies $k \in \{0, 1, ..., N-1\}$ as

$$(M_k f)(n) := \sqrt{N} f(n) U_k(n) \tag{17}$$

This formulation is analogous in construction to classical modulation, defined by pointwise multiplication with a pure harmonic – a Laplacian eigenvector in our case. Classical modulation translates signals in the

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fourier domain; because of the discrete nature of the graph Fourier domain, this property is only weakly shared between the two operators. Instead, the generalized modulation M_k translates the *DC component* of f, $\hat{f}(0)$, to λ_k, i.e. $\widehat{(M_k f)}(\lambda_k) = \hat{f}(0)$. Furthermore, for any function f whose frequency content is localized around λ_0, $(M_k f)$ is localized in frequency around λ_k. Shuman et al. [24] details this construction and provides bounds on spectral localization and other properties.

With these two operators, a graph windowed Fourier atom is constructed[24] for any window function $g \in \mathbb{R}^N$

$$g_{i,k}(n) := (M_k T_i g)(n) = N U_k(n) \sum_{\ell=0}^{N-1} \hat{g}(\lambda_\ell) U_\ell^*(i) U_\ell(n). \tag{18}$$

We can then build a spectrogram $Q = (q_{ik}) \in \mathbb{R}^{N \times N}$ by taking the inner product of each $g_{i,k} \forall i \in \{1, 2, ..., N\} \wedge \forall k \in \{0, 1, ..., N-1\}$ with the target signal f

$$q_{ik} = Sf(i,k) := \langle f, g_{i,k} \rangle. \tag{19}$$

As with the classical windowed Fourier transform, one could interpret this as segmenting the signal by windows and then taking the Fourier transform of each segment

$$q_i = \langle (T_i g \odot f), U \rangle \tag{20}$$

where $\odot$ is the element-wise product.

4.2.3 Using heat kernels of increasing scales to produce the WGFT of the RES

To generate the spectrogram for clustering, we first need a suitable window function. We use the normalized heat kernel as proposed by Shuman et al. [24]

$$\hat{g}(\lambda) = C e^{-t\lambda}, \tag{21}$$

$$C = ||g||_2^{-1}. \tag{22}$$

By translating this kernel, element-wise multiplying it with our target signal f and taking the Fourier transform of the result, we obtain a windowed graph Fourier transform of f that is localized based on the *diffusion distance* [24, 55] from each vertex to every other vertex in the graph.

For an input RES $\mathbf{f}$, signal-biased spectral clustering as proposed by Shuman et al. [24] proceeds as follows:

1. Generate the window matrix P_t, which contains as its columns translated and normalized heat kernels at the scale t
2. Column-wise multiply $F_t = P \odot \mathbf{f}$; the i-th column of F_t is an entry-wise product of the i-th window and $\mathbf{f}$.
3. Take the Fourier Transform of each column of F_t. This matrix, $\hat{C}_t$ is the normalized WGFT matrix.

This produces a single WGFT for the scale t. At this stage, Shuman et al. [24] proposed to saturate the elements of $\hat{C}_t$ using the activation function $\tanh(|\hat{C}_t|)$ (where $|\,.\,|$ is an element-wise absolute value). Then, k-means is performed on this saturated output to yield clusters. This operation has connections to spectral clustering as the features that k-means is run on are coefficients of graph harmonics.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We build upon this approach to add robustness, sensitivity to sign changes, and scalability. Particularly, vertex-frequency clustering builds a set of activated spectrograms at different window scales. These scales are given by simulated heat diffusion over the graph by adjusting the time-scale t in Eqn. 21. Then, the entire set is combined through summation.

4.2.4 Combining the EES and WGFT of the RES

As discussed in Section 2.4, it is useful to consider the sign of the EES in addition to the frequency content of the RES. This is because if we consider two populations of cells, one of which is highly enriched in the experimental condition and another that is enriched in the control, we expect to find similar frequency content of the RES. Namely, both should have very low-frequency content, as indicated in the cartoon in **Fig. 2a**. However, we expect these two populations to have very different EES values. To allow us to distinguish between these populations, we also include the EES in the matrix used for clustering.

We concatenate the EES as an additional column to the multi-resolution spectrogram Q. However, we want to be able to tune the clustering with respect to how much the EES affects the result comapred to the frequency information in Q. Therefore, inspired by spectral clustering as proposed by [69], we first perform PCA on Q to get $k+1$ principle components and then normalize the EES by the $L2$-norm of the first principle component. We then add the EES as an additional column to the PCA-reduced Q to produce the matrix $\hat{Q}$. The weight of the EES can be modulated by a user-adjustable parameter w, but for all experiments in this paper, we leave $w = 1$. Finally, $\hat{Q}$ is used as input for k-means clustering.

The multiscale approach we have proposed has a number of benefits. Foremost, it removes the complexity of picking a window-size. Second, using the actual input signal as a feature allows the clustering to consider both frequency and sign information in the raw experimental signal. For scalability, we leverage the fact that P_t is effectively a diffusion operator and thus can be built efficiently by treating it as a Markov matrix and normalizing the graph adjacency by the degree.

4.2.5 Summary of the vertex frequency clustering algorithm

To identify clusters of cells that are transcriptionally similar and also affected by an experimental perturbation in the same way, we introduced an algorithm called vertex frequency clustering. Our approach builds on previous work by Shuman et al. [24] analyzing the local frequency content of the RES (raw experimental signal) as defined over the vertices of a graph. Here, we introduce two novel adaptations of the algorithm. First, we take a multiresolution approach to quantifying frequency trends in the neighborhoods around each node. By considering windowed signals that are large (i.e. contain many neighboring points) and small (i.e. very proximal on the graph), we can identify clusters both large and small that are similarly affected by an experimental perturbation. Our second contribution is the inclusion of the EES in our basis for clustering. This allows VFC to take into account the degree of enrichment of each group of cells between condition.

Vertex Frequency Clustering is implemented in Python 3 as part of the MELD package and leverages the `graphtools` and `pygsp` packages. MELD is available on GitHub at `https://github.com/KrishnaswamyLab/MELD` and on `pip` as `meld`.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 Methods

5.1 Processing and analysis of the T-cell datasets

Gene expression counts matrices prepared by Datlinger et al. [15] were accessed from the NCBI GEO database accession GSE92872. 3,143 stimulated and 2,597 unstimulated T-cells were processed in a pipeline derived from the published supplementary software. First, artificial genes corresponding to gRNAs were removed from the counts matrix. Genes observed in fewer than five cells were removed. Cell with a library size higher than 35,000 UMI / cell were removed. To filter dead or dying cells, expression of all mitochondrial genes was z-scored and cells with average z-score expression greater than 1 were removed. As in the published analysis, all mitochondrial and ribosomal genes were excluded. Filtered cells and genes were library size normalized and square-root transformed. To impute gene expression, MAGIC was run using default parameters. To build a cell-state graph, 100 PCA dimensions were calculated and edge weights between cells were calculated using an alpha-decay kernel as implemented in the Graphtools library (www.github.com/KrishnaswamyLab/graphtools) using knn=10 and decay=20. To infer the EES, MELD was run on the cell state graph using the stimulated / unstimulated labels and input with the smoothing parameter β = 1. To identify genes that vary with the MELD vector, kNN-DREMI [4] scores were calculated between each gene and the EES vector using default parameters as implemented in scprep (www.github.com/KrishnaswamyLab/scprep). GO term enrichment was performed using EnrichR with the genes having the top 1% of kNN-DREMI scores used as input.

5.2 Processing and analysis of the chordin datasets

Gene expression counts matrices prepared by Wagner et al. [17] (the chordin dataset) were downloaded from NCBI GEO (GSE112294). 16079 cells from *chd* embryos injected with gRNAs targeting chordin and 10782 cells from *tyr* embryos injected with gRNAs targeting tyrosinase were accessed. Lowly expressed genes detected in fewer than 5 cells were removed. Cells with library sizes larger than 15000 UMI / cell were removed. Counts were library-size normalized and square root transformed. Cluster labels included with the counts matrices were used for cell type identification.

During preliminary analysis, a group of 24 cells were identified originating exclusively from the *chd* embryos. Despite an average library size in the bottom 12% of cells, these cells exhibited 546-fold, 246-fold, and 1210-fold increased expression of Sh3Tc1, LOC101882117, and LOC101885394 respectively. To the best of our knowledge, the function of these genes in development is not described. These cells were annotated by Wagner et al. [17] as belonging to 7 cell types including the Tailbud – Spinal Cord and Neural – Midbrain. These cells were excluded from further analysis.

To generate a cell state graph, 100 PCA dimensions were calculated from the square root transformed filtered gene expression matrix of both datasets. Edge weights between cells on the graph were calculated using an alpha-decay kernel with parameters knn=10, decay=10. MAGIC was used to impute gene expression values using t=7. MELD was run using the *tyr* or *chd* label as input. To identify subpopulations of the Tailbud - Presomitic Mesoderm cluster, we applied Vertex Frequency Clustering with k=4. Cell types were annotated using sets of marker genes curated by Farrell et al. [18]. Changes in gene expression for the top and bottom 20% of cells by EES values in the four clusters were compared.

5.3 Generation, processing and analysis of the pancreatic islet datasets

Single-cell RNA-sequencing was performed on human β cells from three different islet donors in the presence and absence of IFNγ. The islets were received on three different days. Cells were cultured for 24 hours

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with 25ng/mL IFNγ (R&D Systems) in CMRL 1066 medium (Gibco) and subsequently dissociated into single cells with 0.05% Trypsin EDTA (Gibco). Cells were then stained with FluoZin-3 (Invitrogen) and TMRE (Life Technologies) and sorted using a FACS Aria II (BD). The three samples were pooled for the sequencing. Cells were immediately processed using the 10X Genomics Chromium 3' Single-Cell RNA-sequencing kit at the Yale Center for Genome Analysis. The raw sequencing data was processed using the 10X Genomics Cell Ranger Pipeline.

Data from all three donors was concatenated into a single matrix for analysis. First, cells not expressing insulin, somatostatin, or glucagon were excluded from analysis using donor-specific thresholds. The data was square root transformed and reduced to 100 PCA dimensions. Next, we applied an MNN kernel to create a graph across all three donors with parameters knn=5, decay=30. This graph was then used for PHATE and MAGIC. The EES was calculated using MELD with default parameters. To identify cell types, we performed Vertex Frequency Clustering using k=8. To identify signature genes of IFNγ stimulation, we calculated kNN-DREMI scores for all genes with the EES vector and kept genes with the top 1% of scores. To identify genes that were differentially expressed in the beta - nonresponsive cluster, we calculated the Wasserstein distance (also called Earth Mover's distance) between expression of each gene in the nonresponsive cluster and all other clusters.

5.4 Quantitative comparisons using Splatter

To generate single-cell data for the quantitative comparisons, we used Splatter. Datasets were all generated using the "Paths" mode so that a latent dimension in the data could be used to create the ground truth likelihood that each cell would be observed in the "experimental" condition relative to the "control". We focused on four data geometries: a tree with three branches, a branch and cluster with either end of the branch enriched or depleted and the cluster unaffected, a single branch with a middle section either enriched or depleted, and four clusters with random segments enriched or depleted. To create clusters, a multi-branched tree was created, and all but the tips of the branches were removed. The ground truth experimental signal was created using custom Python scripts taking the "Steps" latent variable from Splatter and randomly selecting a proportion of each branch or cluster between 10% and 80% of the data was enriched or depleted by 25%. These regions were divided into thirds to create a smooth transition between the unaffected regions and the differentially abundant regions. This likelihood ratio was then centered so that, on average, half the cells would be assigned to each condition. The centered ground truth signal was used to parameterize a Bernoulli random variable and assign each cell to the experimental or control conditions and receive and RES value of +1 or -1, respectively. The data and RES were used as input to the respective algorithms.

To quantify the accuracy of the EES to approximate the ground truth likelihood ratio, we compared the EES, kNN-smoothed signal, or graph averaged signal to the ground truth likelihood of observing each cell in either of the two conditions. We used the Pearson's R statistic because we are only interested in the degree to which these estimates approximate the likelihood ratio. Each of the four data geometries was tested 30 times with different random seeds for scRNA-seq simulation and RES generation.

To quantify the accuracy of VFC to detect the regions of the dataset that were enriched, depleted, or unaffected between conditions, we calculated the Adjusted Rand Score between the ground truth regions with enriched, depleted, or unchanged likelihood ratios between conditions. VFC was compared to k-Means, Spectral Clustering, and Phenograph all using default parameters for all algorithms.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019. ISSN 1744-4292. doi: 10.15252/msb.20188746.

[2] Caleb Weinreb, Samuel Wolock, Allon M. Klein, and Bonnie Berger. SPRING: A kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, April 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx792.

[3] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel Burkhardt, William Chen, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing Transitions and Structure for Biological Data Exploration. *bioRxiv*, page 120378, June 2018. doi: 10.1101/120378.

[4] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.05.061.

[5] Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, and Joshua R. Sanes. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30, August 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.07.054.

[6] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, July 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.05.047.

[7] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics (Oxford, England)*, 31(12):1974–1980, June 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv088.

[8] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W. H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, January 2019. ISSN 1546-1696. doi: 10.1038/nbt.4314.

[9] Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, April 2018. ISSN 1546-1696. doi: 10.1038/nbt.4091.

[10] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, May 2018. ISSN 1546-1696. doi: 10.1038/nbt.4096.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[11] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1254257.

[12] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 2015. ISSN 1474-7596. doi: 10.1186/s13059-015-0844-5.

[13] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra-Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennie T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev, and Levi A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, April 2016. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aad0501.

[14] Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, 167(7):1883–1896.e15, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.039.

[15] Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, January 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4177.

[16] Xin Gao, Deqing Hu, Madelaine Gogol, and Hua Li. ClusterMap: Comparing analyses across multiple Single Cell RNA-Seq profiles. *bioRxiv*, page 331330, June 2018. doi: 10.1101/331330.

[17] Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, page eaar4362, April 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar4362.

[18] Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aar3131.

[19] Kevin R. Moon, Jay S. Stanley, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, February 2018. ISSN 2452-3100. doi: 10.1016/j.coisb.2017.12.008.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[20] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013. ISSN 1053-5888. doi: 10.1109/MSP.2012.2235192.

[21] Martin Barron and Jun Li. Identifying and removing the cell-cycle effect from single-cell rna-sequencing data. *Scientific reports*, 6:33892, 2016.

[22] Smita Krishnaswamy, Matthew H. Spitzer, Michael Mingueneau, Sean C Bendall, Oren Litvin, Erica Stone, Dana Pe'er, and Garry P Nolan. Conditional Density-based Analysis of T cell Signaling in Single Cell Data. *Science (New York, N.Y.)*, 346(6213):1250689, November 2014. ISSN 0036-8075. doi: 10.1126/science.1250689.

[23] David van Dijk, Scott Gigante, Kevin Moon, Alexander Strzalkowski, Katie Ferguson, Jess Cardin, Guy Wolf, and Smita Krishnaswamy. Modeling Dynamics of Biological Systems with Deep Generative Neural Networks. *arXiv:1802.03497 [cs, stat]*, February 2018.

[24] David I Shuman, Benjamin Ricaud, and Pierre Vandergheynst. Vertex-frequency analysis on graphs. *Applied and Computational Harmonic Analysis*, 40(2):260–291, March 2016. ISSN 1063-5203. doi: 10.1016/j.acha.2015.02.005.

[25] L. Le Magoarou, R. Gribonval, and N. Tremblay. Approximate Fast Graph Fourier Transforms via Multilayer Sparse Approximations. *IEEE Transactions on Signal and Information Processing over Networks*, 4(2):407–420, June 2018. ISSN 2373-776X. doi: 10.1109/TSIPN.2017.2710619.

[26] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44 (Web Server issue):W90–W97, July 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw377.

[27] L. A. Turka, D. G. Schatz, M. A. Oettinger, J. J. Chun, C. Gorka, K. Lee, W. T. McCormack, and C. B. Thompson. Thymocyte expression of RAG-1 and RAG-2: Termination by T cell receptor cross-linking. *Science*, 253(5021):778–781, August 1991. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1831564.

[28] M. Hammerschmidt, F. Pelegri, M. C. Mullins, D. A. Kane, F. J. van Eeden, M. Granato, M. Brand, M. Furutani-Seiki, P. Haffter, C. P. Heisenberg, Y. J. Jiang, R. N. Kelsh, J. Odenthal, R. M. Warga, and C. Nusslein-Volhard. Dino and mercedes, two genes regulating dorsal development in the zebrafish embryo. *Development*, 123(1):95–102, December 1996. ISSN 0950-1991, 1477-9129.

[29] Stefan Schulte-Merker, Kevin J. Lee, Andrew P. McMahon, and Matthias Hammerschmidt. The zebrafish organizer requires *chordino*. *Nature*, 387(6636):862–863, June 1997. ISSN 1476-4687. doi: 10.1038/43092.

[30] Shannon Fisher and Marnie E. Halpern. Patterning the zebrafish axial skeleton requires early *chordin* function. *Nature Genetics*, 23(4):442–446, December 1999. ISSN 1546-1718. doi: 10.1038/70557.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[31] Anastasia Katsarou, Soffia Gudbjörnsdottir, Araz Rawshani, Dana Dabelea, Ezio Bonifacio, Barbara J. Anderson, Laura M. Jacobsen, Desmond A. Schatz, and Åke Lernmark. Type 1 diabetes mellitus. *Nature Reviews Disease Primers*, 3:17016, March 2017. ISSN 2056-676X. doi: 10.1038/nrdp.2017.16.

[32] V. Ablamunits, D. Elias, T. Reshef, and I. R. Cohen. Islet T cells secreting IFN-gamma in NOD mouse diabetes: Arrest by p277 peptide treatment. *Journal of Autoimmunity*, 11(1):73–81, February 1998. ISSN 0896-8411. doi: 10.1006/jaut.1997.0177.

[33] Andrew S. Diamond and Ronald G. Gill. An Essential Contribution by IFN-γ to CD8+ T Cell-Mediated Rejection of Pancreatic Islet Allografts. *The Journal of Immunology*, 165(1):247–255, July 2000. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.165.1.247.

[34] Miguel Lopes, Burak Kutlu, Michela Miani, Claus H. Bang-Berthelsen, Joachim Størling, Flemming Pociot, Nathan Goodman, Lee Hood, Nils Welsh, Gianluca Bontempi, and Decio L. Eizirik. Temporal profiling of cytokine-induced genes in pancreatic β-cells by meta-analysis and network inference. *Genomics*, 103(4):264–275, April 2014. ISSN 0888-7543. doi: 10.1016/j.ygeno.2013.12.007.

[35] Mauro J. Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A. Engelse, Francoise Carlotti, Eelco J.P. de Koning, and Alexander van Oudenaarden. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3, October 2016. ISSN 2405-4712. doi: 10.1016/j.cels.2016.09.002.

[36] Chilakamarti V Ramana, M. Pilar Gil, Robert D Schreiber, and George R Stark. Stat1-dependent and -independent pathways in IFN-γ-dependent signaling. *Trends in Immunology*, 23(2):96–101, February 2002. ISSN 1471-4906. doi: 10.1016/S1471-4906(01)02118-4.

[37] Anthony J. Sadler and Bryan R. G. Williams. Interferon-inducible antiviral effectors. *Nature reviews. Immunology*, 8(7):559–568, July 2008. ISSN 1474-1733. doi: 10.1038/nri2314.

[38] Katherine A. Fitzgerald. The Interferon Inducible Gene: Viperin. *Journal of Interferon & Cytokine Research*, 31(1):131–135, January 2011. ISSN 1079-9907. doi: 10.1089/jir.2010.0127.

[39] Zhiwei Zheng, Lin Wang, and Jihong Pan. Interferon-stimulated gene 20-kDa protein (ISG20) in infection and disease: Review and outlook. *Intractable & Rare Diseases Research*, 6(1):35–40, February 2017. ISSN 2186-3644. doi: 10.5582/irdr.2017.01004.

[40] Monica Hultcrantz, Michael H. Hühn, Monika Wolf, Annika Olsson, Stella Jacobson, Bryan R. Williams, Olle Korsgren, and Malin Flodström-Tullberg. Interferons induce an antiviral state in human pancreatic islet cells. *Virology*, 367(1):92–101, October 2007. ISSN 0042-6822. doi: 10.1016/j.virol.2007.05.010.

[41] Andrew F. Stewart, Mehboob A. Hussain, Adolfo García-Ocaña, Rupangi C. Vasavada, Anil Bhushan, Ernesto Bernal-Mizrachi, and Rohit N. Kulkarni. Human β-Cell Proliferation and Intracellular Signaling: Part 3. *Diabetes*, 64(6):1872–1885, June 2015. ISSN 0012-1797. doi: 10.2337/db14-1843.

[42] Yurong Xin, Giselle Dominguez Gutierrez, Haruka Okamoto, Jinrang Kim, Ann-Hwee Lee, Christina Adler, Min Ni, George D. Yancopoulos, Andrew J. Murphy, and Jesper Gromada. Pseudotime Ordering of Single Human β-Cells Reveals States of Insulin Production and Unfolded Protein Response. *Diabetes*, 67(9):1783–1794, September 2018. ISSN 0012-1797, 1939-327X. doi: 10.2337/db18-0365.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[43] Lydia Farack, Matan Golan, Adi Egozi, Nili Dezorella, Keren Bahar Halpern, Shani Ben-Moshe, Immacolata Garzilli, Beáta Tóth, Lior Roitman, Valery Krizhanovsky, and Shalev Itzkovitz. Transcriptional Heterogeneity of Beta Cells in the Intact Pancreas. *Developmental Cell*, 48(1):115–125.e4, January 2019. ISSN 1534-5807. doi: 10.1016/j.devcel.2018.11.001.

[44] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, September 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1305-0.

[45] B. van Ginneken and A. Mendrik. Image Denoising with k-nearest Neighbor and Support Vector Regression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 603–606, August 2006. doi: 10.1109/ICPR.2006.685.

[46] Florian Wagner, Yun Yan, and Itai Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*, page 217737, April 2018. doi: 10.1101/217737.

[47] Fan Zhang and Edwin R. Hancock. Graph spectral image smoothing using the heat kernel. *Pattern Recognition*, 41(11):3328–3342, November 2008. ISSN 0031-3203. doi: 10.1016/j.patcog.2008.05.007.

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, October 2011. ISSN 1533-7928.

[49] Y. P Mack and M Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, March 1979. ISSN 0047-259X. doi: 10.1016/0047-259X(79)90065-4.

[50] Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011. ISSN 1935-7524. doi: 10.1214/11-EJS606.

[51] Yi-Hung Kung, Pei-Sheng Lin, and Cheng-Hsiung Kao. An optimal k-nearest neighbor for density estimation. *Statistics & Probability Letters*, 82(10):1786–1791, October 2012. ISSN 0167-7152. doi: 10.1016/j.spl.2012.05.017.

[52] Ulrike Von Luxburg and Morteza Alamgir. Density estimation from unweighted k-nearest neighbor graphs: A roadmap. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 225–233. Curran Associates, Inc., 2013.

[53] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, February 2018. ISBN 978-1-315-14091-9. doi: 10.1201/9781315140919.

[54] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[55] Nathanael Perraudin, Benjamin Ricaud, David Shuman, and Pierre Vandergheynst. Global and local uncertainty principles for signals on graphs. *arXiv preprint arXiv:1603.03030*, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[56] Stephane Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, December 2008. ISBN 978-0-08-092202-7.

[57] Dengyong Zhou and Bernhard Schölkopf. A regularization framework for learning from graph data. In *ICML workshop on statistical relational learning and Its connections to other fields*, volume 15, pages 67–8, 2004.

[58] Jihun Ham, Daniel D Lee, and Lawrence K Saul. Semisupervised alignment of manifolds. In *AISTATS*, pages 120–127, 2005.

[59] Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer, 2004.

[60] Rie K. Ando and Tong Zhang. Learning on Graph with Laplacian Regularization. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 25–32. MIT Press, 2007.

[61] Kilian Q Weinberger, Fei Sha, Qihui Zhu, and Lawrence K. Saul. Graph Laplacian Regularization for Large-Scale Semidefinite Programming. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1489–1496. MIT Press, 2007.

[62] X. He, M. Ji, C. Zhang, and H. Bao. A Variance Minimization Criterion to Feature Selection Using Laplacian Regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10): 2013–2025, October 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.44.

[63] X. Liu, D. Zhai, D. Zhao, G. Zhai, and W. Gao. Progressive Image Denoising Through Hybrid Graph Laplacian Regularization: A Unified Framework. *IEEE Transactions on Image Processing*, 23(4): 1491–1503, April 2014. ISSN 1057-7149. doi: 10.1109/TIP.2014.2303638.

[64] J. Pang, G. Cheung, A. Ortega, and O. C. Au. Optimal graph laplacian regularization for natural image denoising. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2294–2298, April 2015. doi: 10.1109/ICASSP.2015.7178380.

[65] Jiahao Pang and Gene Cheung. Graph Laplacian Regularization for Image Denoising: Analysis in the Continuous Domain. *IEEE Transactions on Image Processing*, 26(4):1770–1785, April 2017. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2017.2651400.

[66] Nathanaël Perraudin, Johan Paratte, David Shuman, Lionel Martin, Vassilis Kalofolias, Pierre Vandergheynst, and David K. Hammond. GSPBOX: A toolbox for signal processing on graphs. *ArXiv e-prints*, August 2014.

[67] David I Shuman, Pierre Vandergheynst, and Pascal Frossard. Chebyshev polynomial approximation for distributed signal processing. In *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, pages 1–8. IEEE, 2011.

[68] Nathanaël Perraudin, Nicki Holighaus, Peter L Søndergaard, and Peter Balazs. Designing Gabor windows using convex optimization. *arXiv preprint arXiv:1401.6033*, 2014.

[69] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

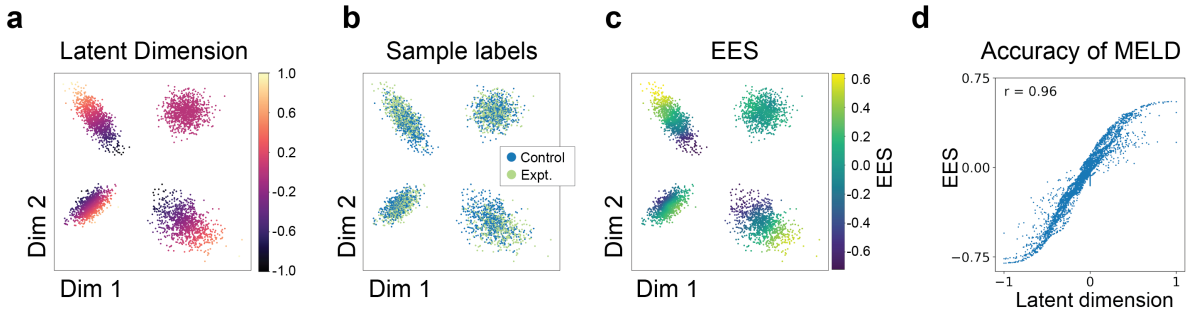

Figure S1: MELD captures the latent experimental signals across clusters. (**a**) In many scRNA-seq experiments, there is not one, but many populations of cells. Each of these populations, or cell types, may respond to an experimental perturbation differently. We simulated four Gaussian clouds of various sizes and densities and created artificial latent dimensions across each population. The scale of this dimension is arbitrarily defined over the interval [-1,1], Note that the axis of greatest variation within a population does not always match the dimension corresponding to the experimental response, as in the lower left cluster. Furthermore, some populations of cells may not respond to the experimental perturbation, as in the upper right cluster. (**b**) To simulate the results of noisy experimental sampling of these cell populations, we assigned experimental labels to cells such that cells with high latent dimension values are more likely to come from the experimental condition and cells with low latent dimension values are likely to come from the control experiment. These labels are used as the Raw Experimental Signal (RES). (**c**) MELD identifies an Enhanced Experimental Signal (EES) in each cluster. (**d**) Comparing the EES to the ground truth Latent Dimension, we find very strong correlation between the EES and the true experimental signal.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

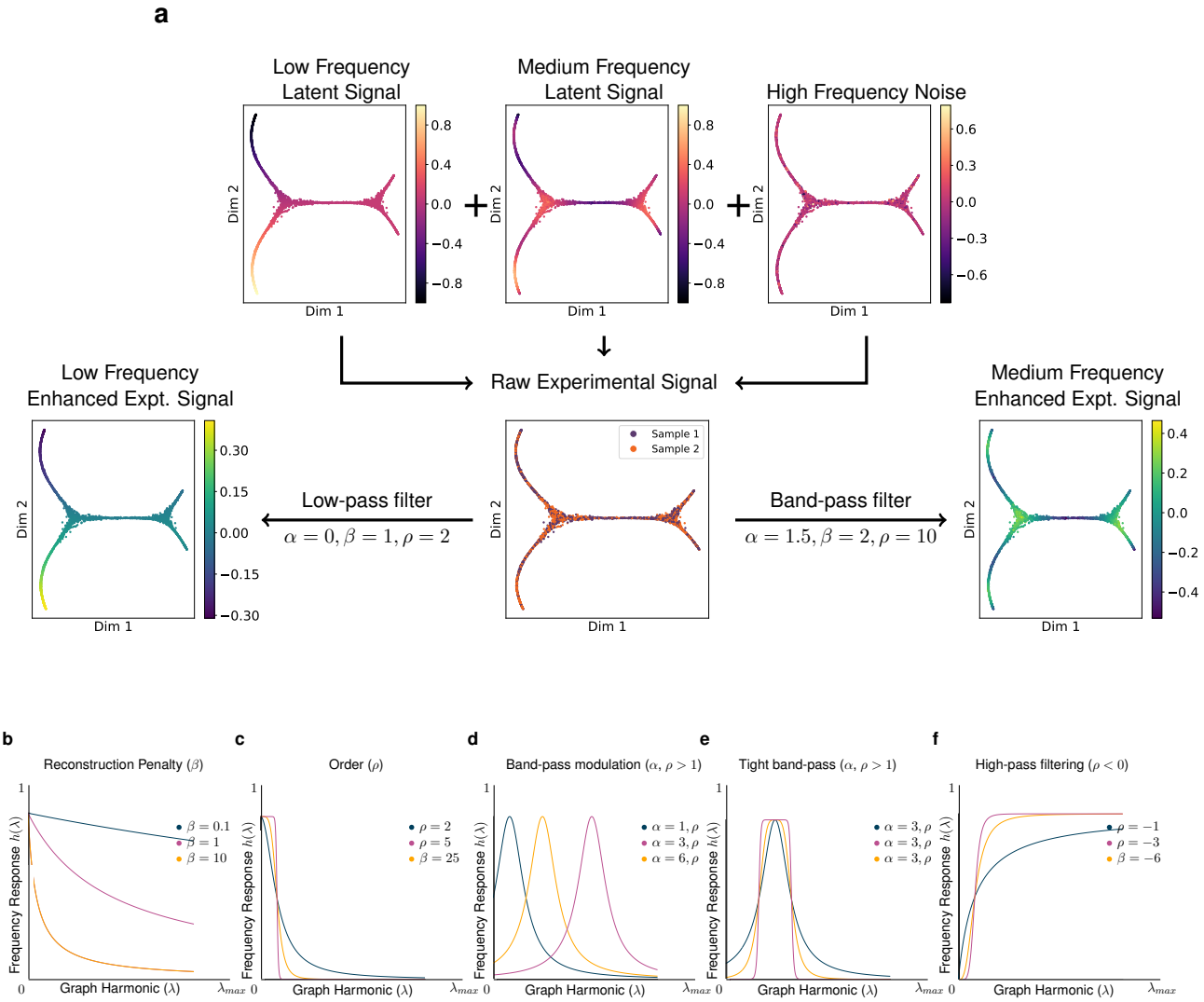

Figure S2: Source Separation and Parameter Analysis with the MELD filter. (**a**) A raw experimental signal (center) is obtained that is a binarized observation of a low frequency latent signal (top left), a medium frequency latent signal (top middle), and high frequency noise (top right). Analysis of the RES alone is intractable as it is corrupted by noise and experimental binarization. MELD low-pass filters (bottom left) to separate a longitudinal trajectory and band-pass filters (bottom right) to yield the periodic signature of the medium frequency latent signal. Parameters used for this analysis are supplied beneath the corresponding arrows. (**b**) Reconstruction penalty β controls a low-pass filter. For this demonstration, $\alpha = 0, \rho = 1$. This filter is equivalent to Laplacian regularization. (**c**) Order ρ controls the filter squareness. This parameter is used in the low-pass filter of (**a**). For this demonstration, $\beta = 1, \alpha = 0$. (**d**) Band-pass modulation via α. When ρ is even valued, α modulates the central frequency of a band-pass filter. This parameter is used in (**a**) to separate a medium-frequency source from a low-frequency source. (**e**) α and ρ combine to make square band-pass filters. For (**d**) and (**e**), $\beta = 1$. (**f**) Negative values of ρ yield a high-pass filter. For (**b-f**), Laplacian harmonics for a general normalized Laplacian are plotted on the x-axis. The frequency response of the filter given by the colored parameters is on the y-axis.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

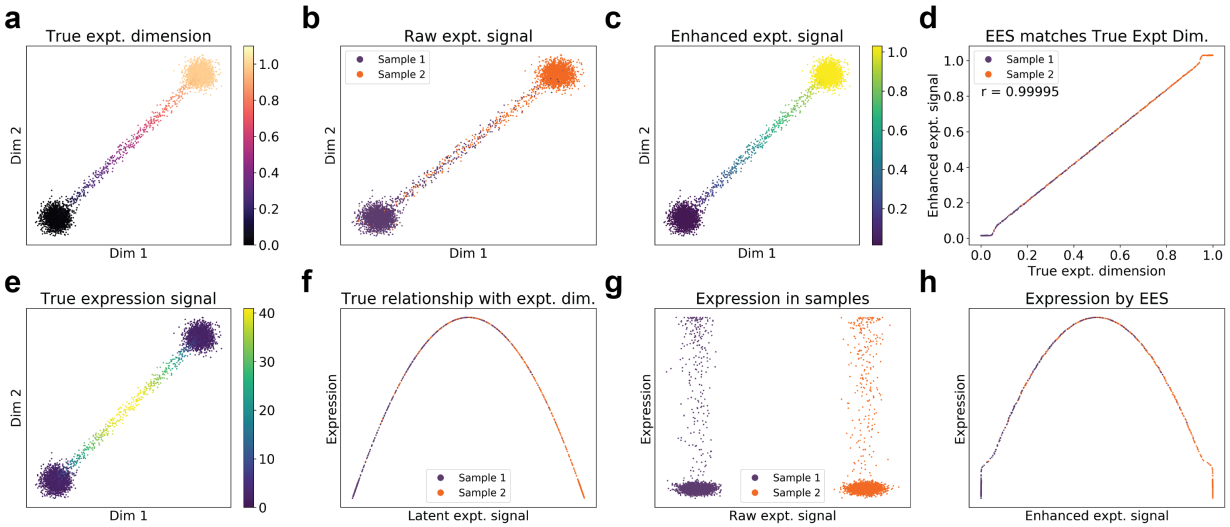

Figure S3: MELD can capture a ground-truth non-linear gene expression signature. (**a**) To demonstrate the ability for MELD to capture non-linear, and non-monotonic gene expression signatures, we simulated a simple 100-dimensional datatset with two terminal cell states connected by an intermediate, transitional spectrum of cells with added noise. The true latent experimental dimension (corresponding to the progression instigated by an experimental condition) is a smooth progression from the left to the right terminal cell state. (**b**) To simulate the results of noisy experimental sampling of these cell populations, we assigned experimental labels to cells such that cells with high latent dimension values are more likely to come from the experimental condition and cells with low latent dimension values are likely to come from the control experiment. These labels are used as the Raw Experimental Signal (RES). (**c**) MELD identifies an Enhanced Experimental Signal (EES). (**d**) Comparing the EES to the ground truth experimental dimension, we find very strong recovery of the true experimental signal. (**e**) To simulate a non-linear gene expression pattern of a single gene, we created an artificial gene expression signal that is low in the terminal cell states, but peaks in the intermediate transitional cells. (**f**) Plotting the expression of the artifical gene as a function of the true experimental dimension, we can observe the non-linear nature of the artificial expression signal. (**g**) Using only the sample labels to characterize expression of this gene in our simulated dataset, we observe no difference in expression of the gene between conditions. (**h**) Only when plotting the expression as a function of the enhance experimental signal can we observe the non-linear nature of the expression. This would be hidden without MELD.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

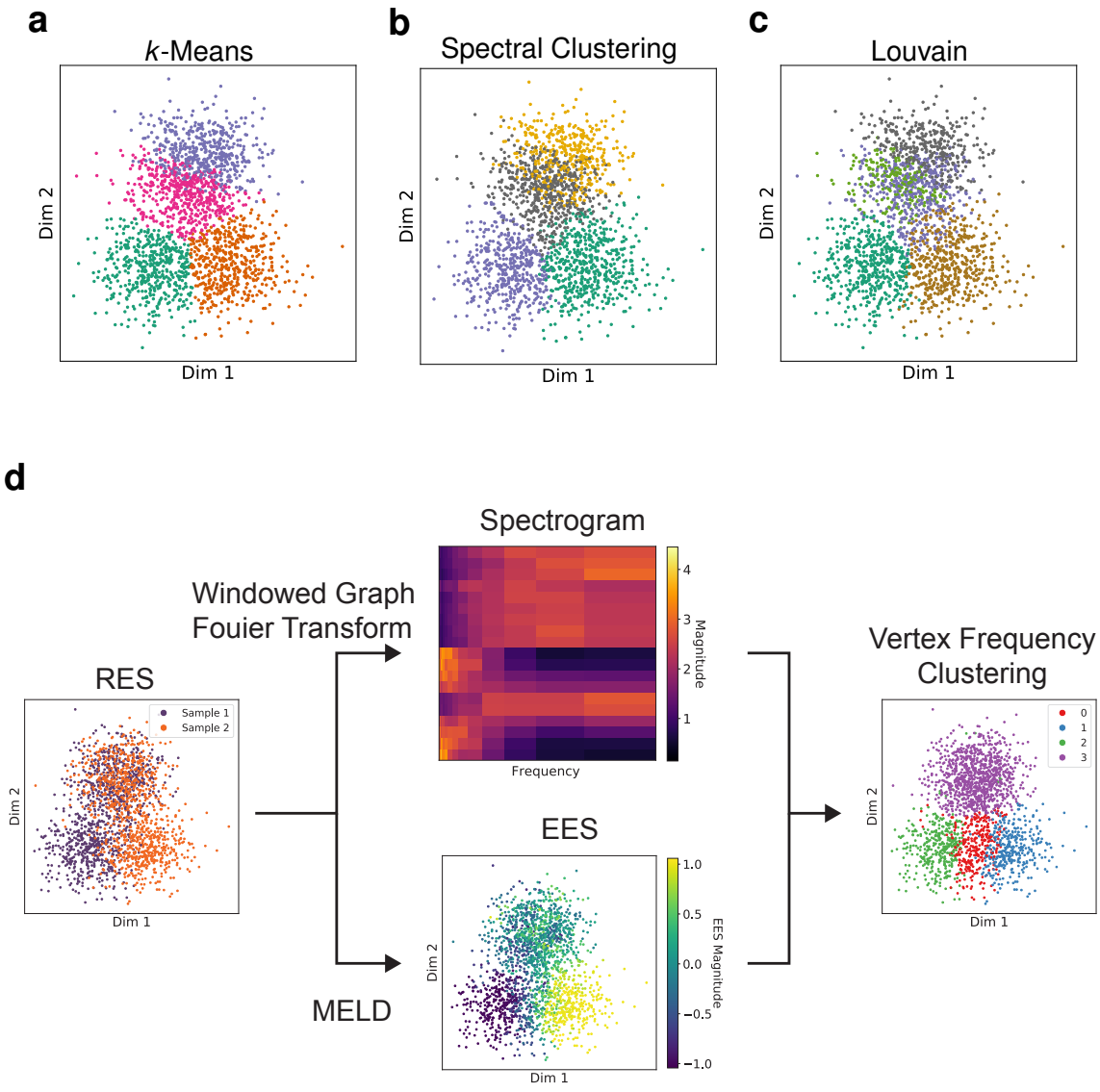

Figure S4: Vertex-Frequency clustering with MELD. A Gaussian mixture model was used to generate N = 1000 points in a mixture of three Gaussian distributions. This experiment is representative of a two-cell type experiment (split by Dim 2) in which one sample changes (bottom clusters) along Dim 1 due to the experiment while the other remains mixed (top clusters). (**a**) *k*-Means clustering separates the left and right experimental groups but splits the upper group erroneously. (**b**) Spectral clustering replicates the performance of *k*-Means in this example. (**c**) Louvain modularity clustering splits the mixture into five groups, with the same lower separations as before but with three groups in the upper cell type. **(d)** Vertex-Frequency clustering recovers a new cluster type. Briefly, the RES (left) is used for (1) a windowed graph Fourier Transform to obtain vertex-frequency information (above, logarithmically downsampled for clarity) and (2) MELD, which generates a continuous profile of the simulated experimental effect. These measures are concatenated together and clustered with *k*-Means. The clusters (right) separate the two cell types (purple and green/red/blue), and finds a separate grouping of cells that are in transition from green to blue, shown in red. One may see that in the spectrogram the green and blue groups are found on relatively low frequency patterns (bottom half of spectrogram, mostly black bands), whereas the medium frequency transition is well separated (middle of bottom bands). The well-mixed, nonresponsive population is entirely high frequency (top half).

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

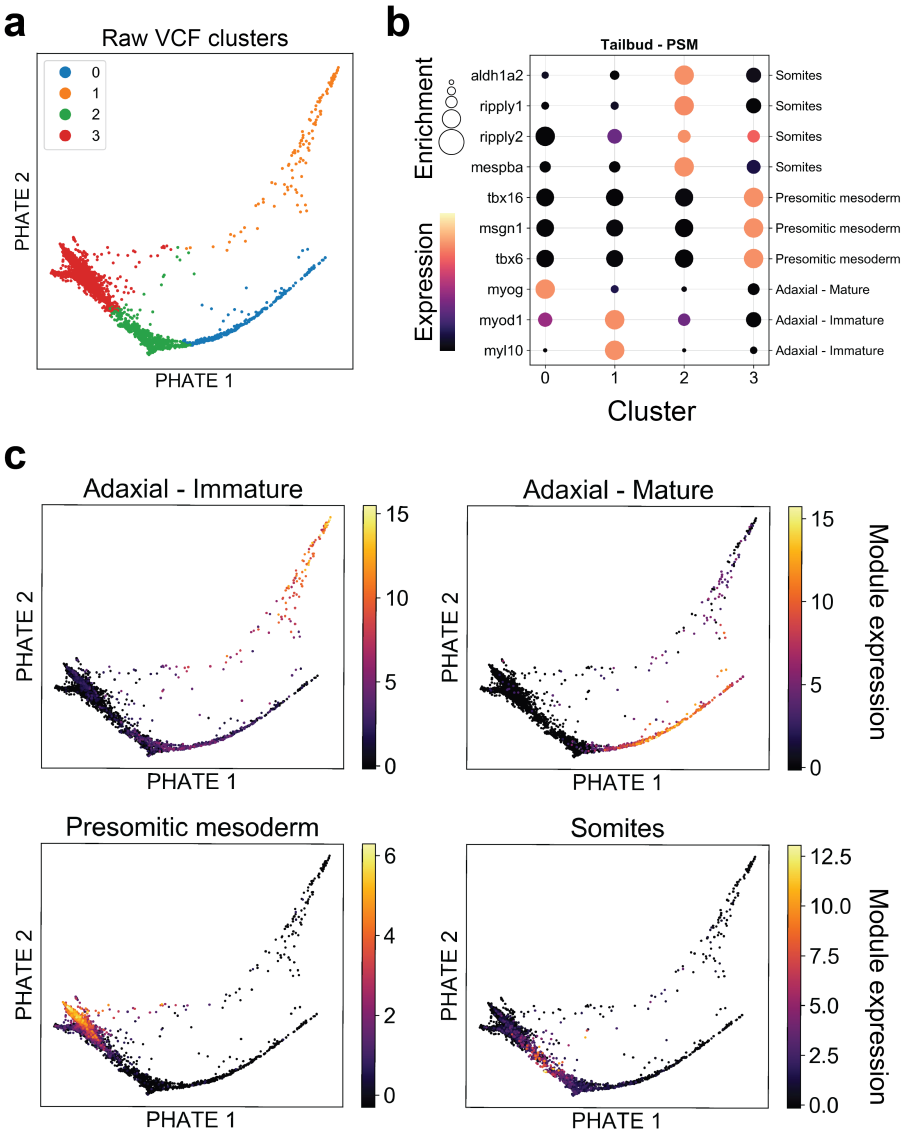

Figure S5: Characterization of vertex-frequency clusters in the Tailbud - Presomitic Mesoderm Cluster (**a**) Raw vertex-frequency cluster assignments on a PHATE visualization. (**b**) Normalized expression of previously identified marker genes of possible subtypes of the Tailbud - Presomitic Mesoderm[18]. The color of the dot for each gene in each cluster indicates the expression level after MAGIC and the size of the dot corresponds to the normalized Wasserstein distance between expression within cluster to all other clusters. (**c**) Average z-score transformed expression of genes associated with each cell type is plotted on a PHATE visualization of the Tailbud - Presomitic Mesoderm Cluster.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

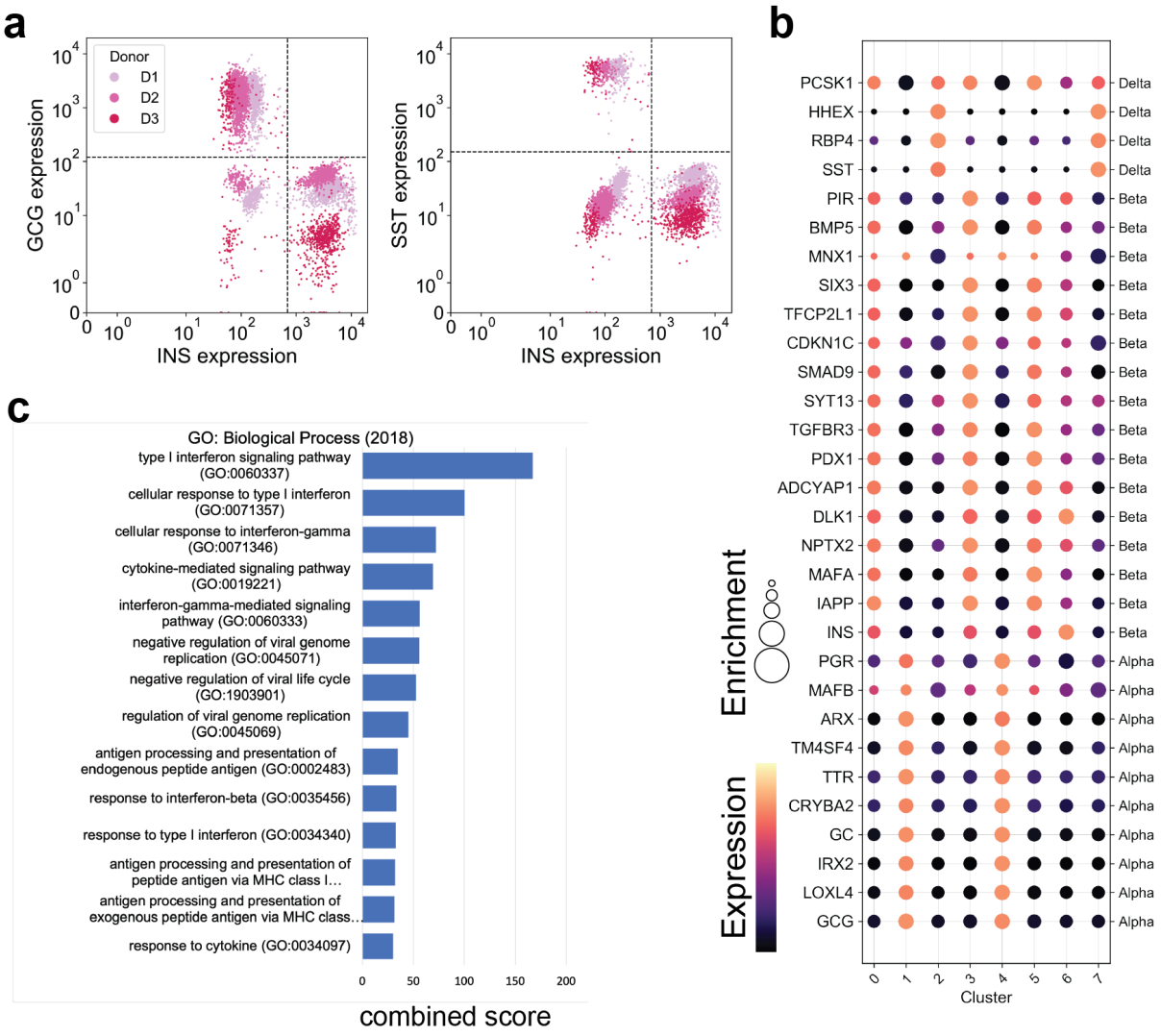

Figure S6: Analysis of pancreatic islet cells from three donors. (**a**) Library-size normalized expression of insulin (INS), glucagon (GCG), and somatostatin (SST) shows donor-specific batch effect across islet cells. (**b**) Normalized expression of previously identified marker genes of alpha, beta, and delta cells[35] in each cluster. The color of the dot for each gene in each cluster indicates the expression level after MAGIC and the size of the dot corresponds to the normalized Wasserstein distance between expression within cluster to all other clusters. (**c**) Results of Enrichr[26] gene set enrichment of 491 signature genes identified in at least one cell-type shows strong enrichment for genes in the interferon signalling pathways.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

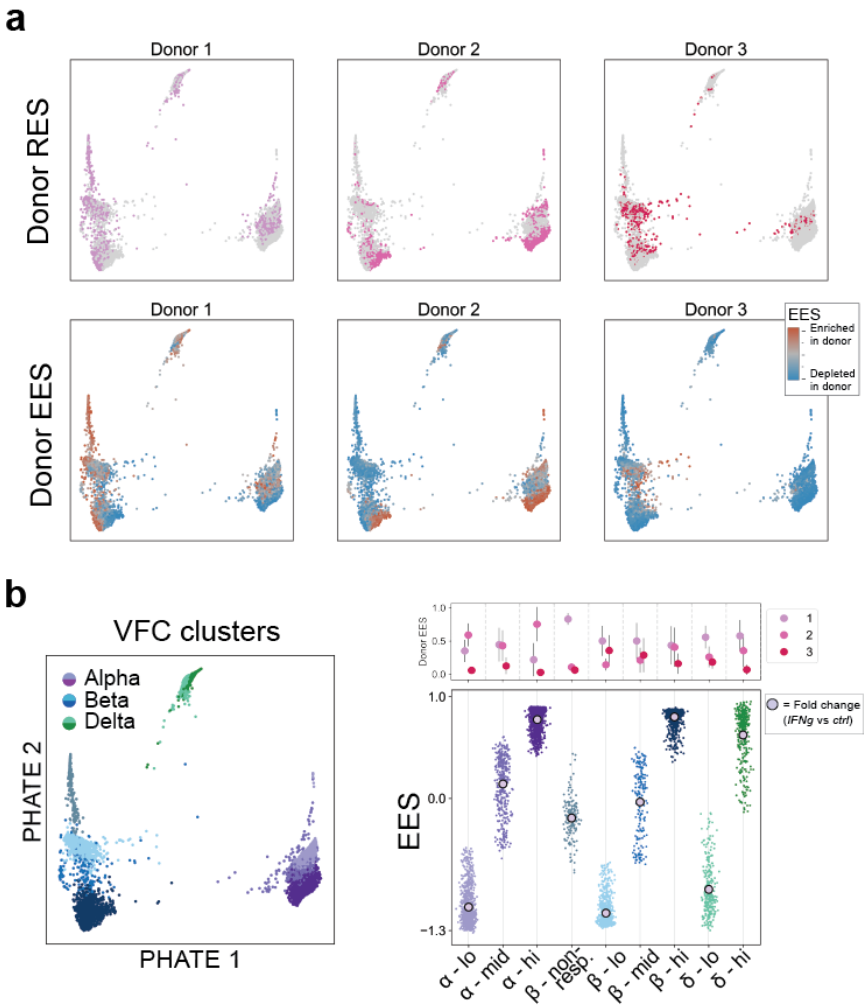

Figure S7: Analysis of islet cell profiles across donors. (**a**) The RES and EES associated with each donor from which islet cells were obtained. (**b**) Comparison of the EES values within each vertex frequency cluster identifies changes in enrichment for each cluster in various donors.

bioRxiv preprint doi: https://doi.org/10.1101/532846; this version posted September 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

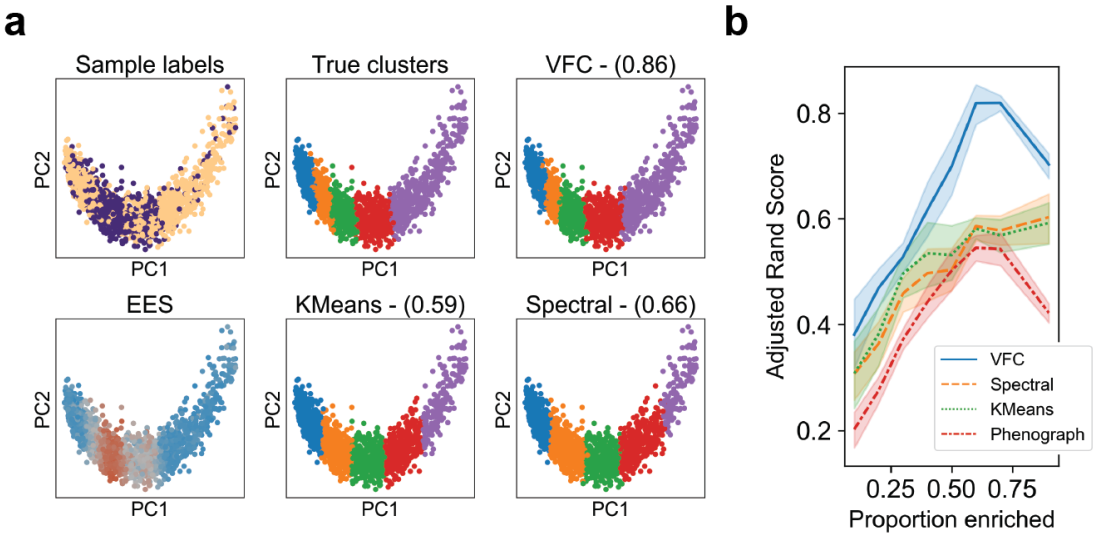

Figure S8: VFC comparisons on the non-monotonic branch data geometry. (**a**) The sample labels, EES, and clustering results from one representative simulation. (**b**) The Adjusted Rand Score of each method as a function of the region of the branch that is enriched. This region is expressed as a proportion of the branch that is either enriched or depleted. This region is evenly divided into thirds to create the ground truth signal for EES and VFC quantification. We observe that the large spread in scores in this case is related to the interval width. Across large ranges of enriched region sizes, VFC outperforms comparison methods. Phenograph was not included in this analysis because of it performed the worst in the cross-dataset comparison.