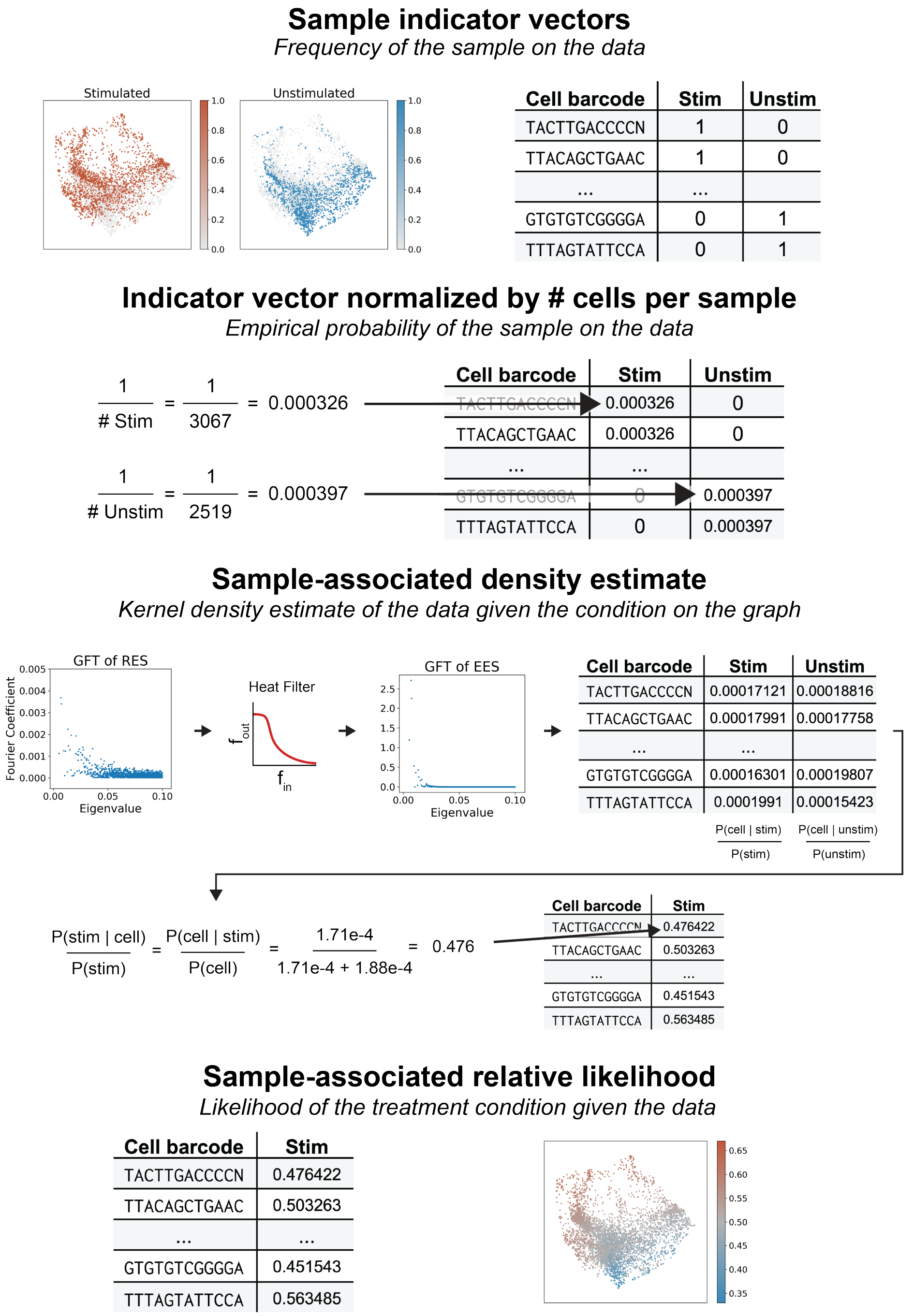
natureresearch

https://doi.org/10.1038/s41587-020-00803-5

Supplementary information

Quantifying the effect of experimental perturbations at single-cell resolution

In the format provided by the authors and unedited



Sample indicator vectors

Frequency of the sample on the data

Cell barcode	Stim	Unstim
TACTTGACCCCN	1	0
TTACAGCTGAAC	1	0
...	...	
GTGTGTCGGGGA	0	1
TTTAGTATTCCA	0	1

Indicator vector normalized by # cells per sample

Empirical probability of the sample on the data

$$\frac{1}{\#\ \text{Stim}} = \frac{1}{3067} = 0.000326$$

$$\frac{1}{\#\ \text{Unstim}} = \frac{1}{2519} = 0.000397$$

Cell barcode	Stim	Unstim
TACTTGACCCCN	0.000326	0
TTACAGCTGAAC	0.000326	0
...	...	
GTGTGTCGGGGA	0	0.000397
TTTAGTATTCCA	0	0.000397

Sample-associated density estimate

Kernel density estimate of the data given the condition on the graph

Cell barcode	Stim	Unstim
TACTTGACCCCN	0.00017121	0.00018816
TTACAGCTGAAC	0.00017991	0.00017758
...	...	
GTGTGTCGGGGA	0.00016301	0.00019807
TTTAGTATTCCA	0.0001991	0.00015423
	$\frac{P(\text{cell} \mid \text{stim})}{P(\text{stim})}$	$\frac{P(\text{cell} \mid \text{unstim})}{P(\text{unstim})}$

$$\frac{P(\text{stim} \mid \text{cell})}{P(\text{stim})} = \frac{P(\text{cell} \mid \text{stim})}{P(\text{cell})} = \frac{1.71e\text{-}4}{1.71e\text{-}4 + 1.88e\text{-}4} = 0.476$$

Cell barcode	Stim
TACTTGACCCCN	0.476422
TTACAGCTGAAC	0.503263
...	...
GTGTGTCGGGGA	0.451543
TTTAGTATTCCA	0.563485

Sample-associated relative likelihood

Likelihood of the treatment condition given the data

Cell barcode	Stim
TACTTGACCCCN	0.476422
TTACAGCTGAAC	0.503263
...	...
GTGTGTCGGGGA	0.451543
TTTAGTATTCCA	0.563485

Figure S1: A step-by-step visual representation of the sample-associated relative likelihood algorithm using data from Datlinger et al. [1]. The sample labels are used to create a one-hot indicator signal for each condition. These one-hot signals are then column-wise L1-normalized such that the sum of each vector is 1. This gives each sample equal weight over the manifold despite a potential uneven number of cells in each condition. Next, the manifold heat filter is used to calculate a kernel density estimate for each condition. These sample-associated density estimates are then row-wise L1-normalized to yield the relative likelihood that each cell would be observed in each condition. The relative likelihood of the treatment condition relative to the control is used for two-condition experiments.

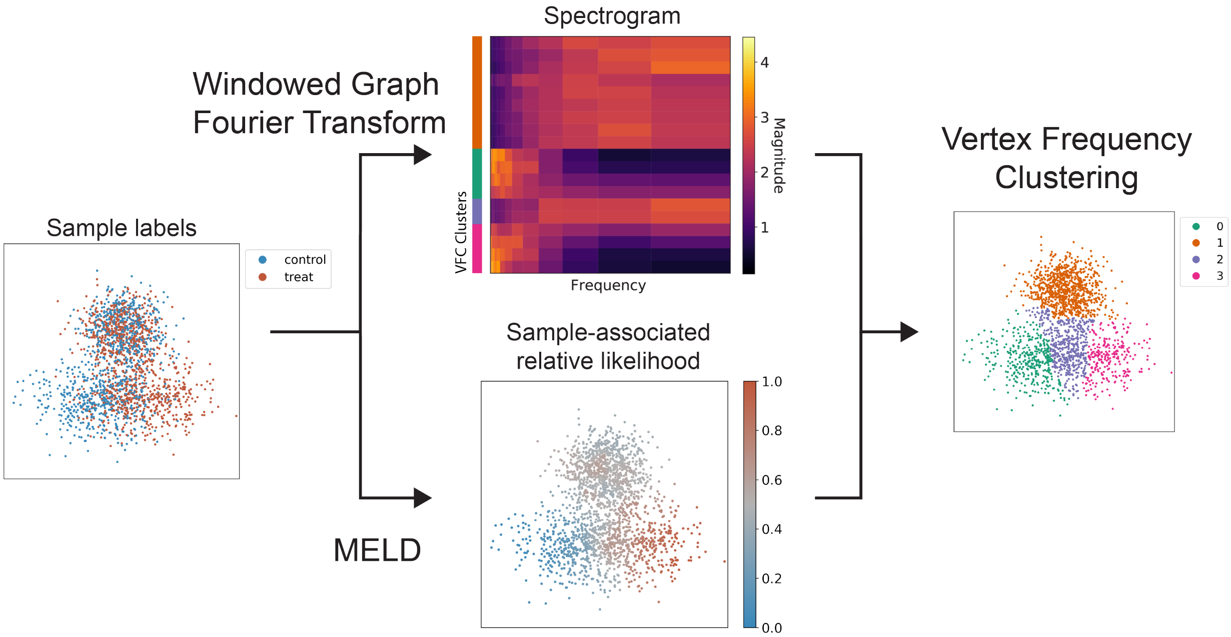


Figure S2: Vertex-Frequency clustering with MELD. A Gaussian mixture model was used to generate N = 2000 points in a mixture of three Gaussian distributions. This experiment is representative of a two-cell type experiment (split by Dim 2) in which one sample changes (bottom clusters) along Dim 1 due to the experiment while the other remains mixed (top clusters). Briefly, the sample labels (left) are used for (1) a windowed graph Fourier Transform to obtain vertex-frequency information (above, logarithmically downsampled for clarity) and (2) to calculate the sample-associated relative likelihood. These measures are concatenated together and clustered with *k*-Means. The clusters (right) separate the two groups of data (orange and green/purple/pink), and finds a separate grouping of points that are in transition from green to pink, shown in purple. One may see along the left side of the spectrogram that points in the green and pink clusters are found on relatively low frequency patterns with high activations in lower frequencies, whereas the transition group in purple has a well-separated medium frequency pattern. The well-mixed, nonresponsive population is entirely high frequency.

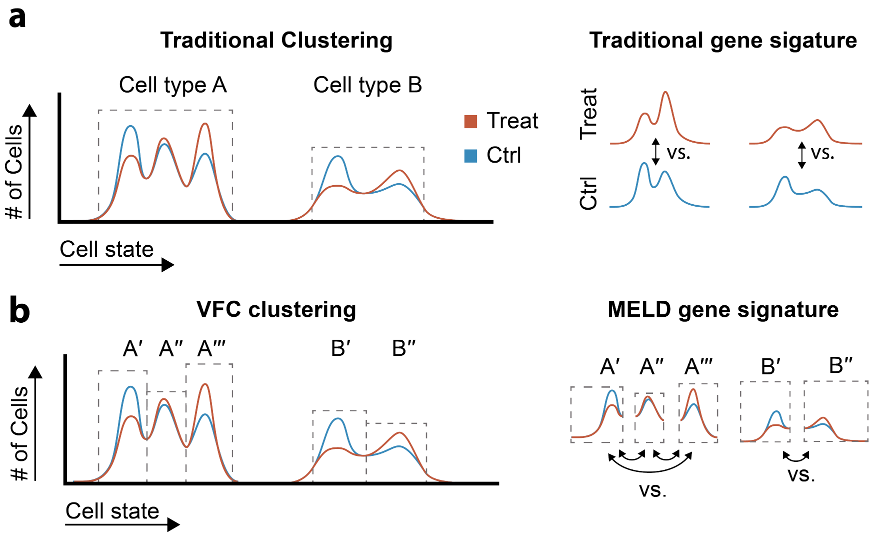
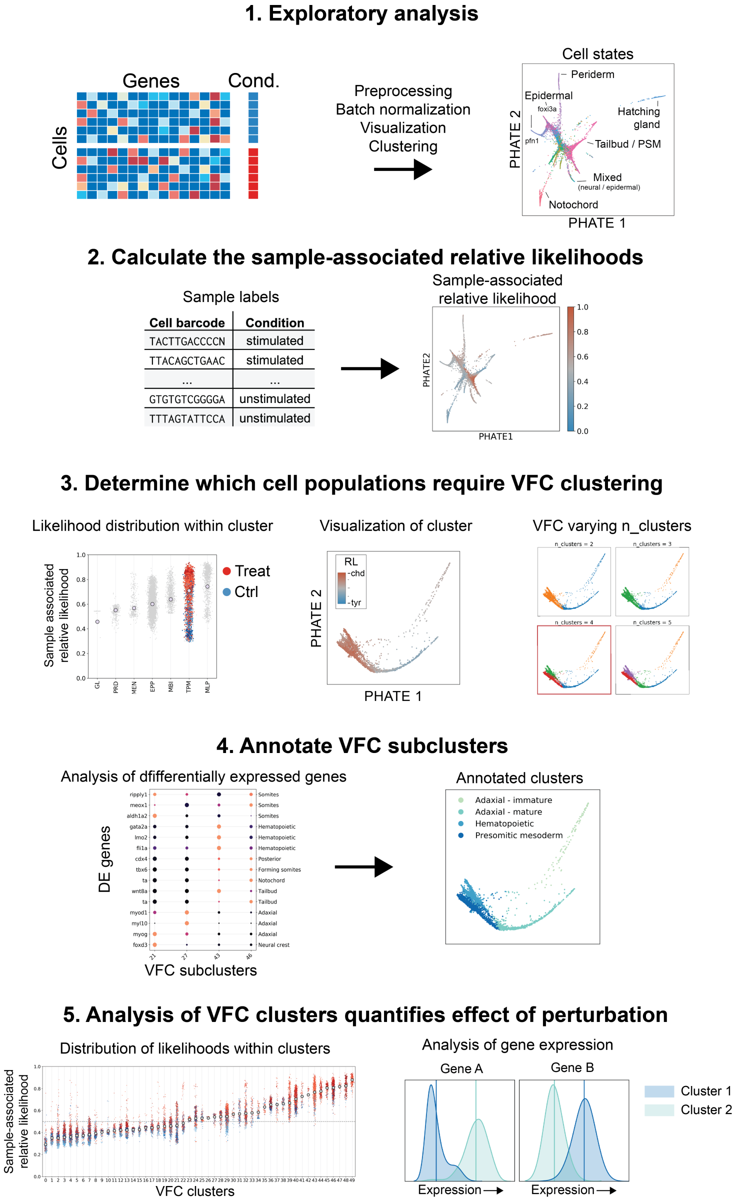


Figure S3: Identifying gene signatures using MELD. (**a**) In traditional gene signature analysis, clusters are identified based on data geometry and may not capture subpopulations of cells with varying response to a perturbation. In this framework, gene signatures are calculated by comparing cells from the experimental and control condition within each cluster. (**b**) To identify gene signatures of a perturbation with the MELD toolkit, we propose first partitioning cell populations with divergent responses to an experimental perturbation prior to differential expression analysis. We then assume that the differences within each VFC cluster is noise. Differential expression can either be calculated between subclusters identified by VFC (as shown) or by comparing each VFC cluster to the rest of the dataset independently.



Cell barcode	Condition
TACTTGACCCCN	stimulated
TTACAGCTGAAC	stimulated
...	...
GTGTGTCGGGGA	unstimulated
TTTAGTATTCCA	unstimulated

Figure S4: Overview of a pipeline for single cell analysis using MELD. (**1.**) Initial exploratory analysis of the dataset should follow established best practices to identify coarse-grained cell populations [2, 3]. (**2.**) Calculating the sample-associated relative likelihood provides a measure for each cell describing the probability that cell would be observed in the experimental condition relative to the control. (**3.**) To identify populations most affected by a perturbation, we consider several sources of information regarding biological heterogeneity and the effect of the perturbation within each exploratory cluster. We then apply VFC at the determined cluster resolution. (**4.**) To assess the biological relevance of each VFC cluster, standard methods for cluster annotation can be applied. (**5.**) To characterize the gene signature of the perturbation, we compare expression differences between VFC clusters with varying relative likelihood distributions.

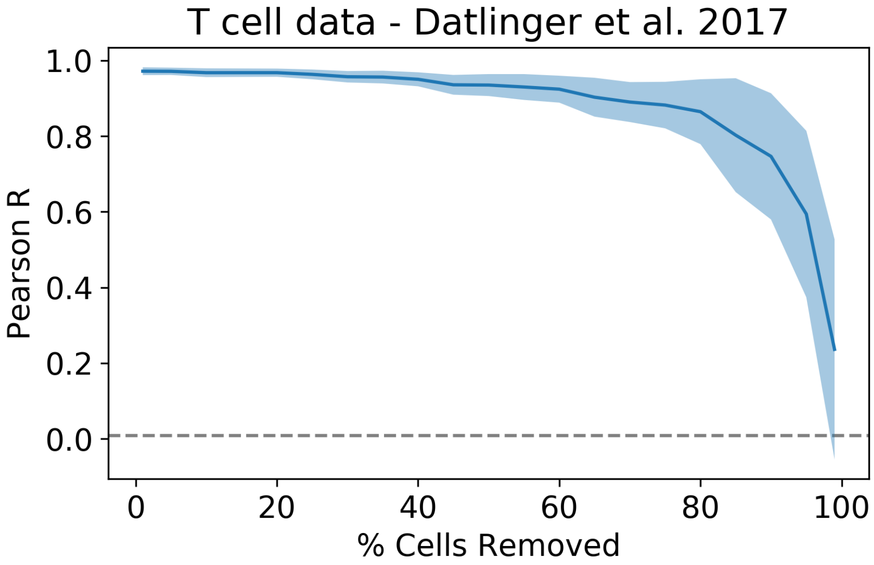


Figure S5: Result of down-sampling on accurately recovering simulated relative likelihood values. We generated 100 random ground truth relative likelihoods and then removed between 1-99% of the cells in the dataset before running the MELD with default parameters. The average Pearson's R is shown as a function of the number of cells removed prior to estimating the sample-associated relative likelihood. The shaded area demarks ±1 standard deviation. We observe an average correlation >0.9 for all experiments with at least 35% of the data present, or 1956 out of 5591 cells.

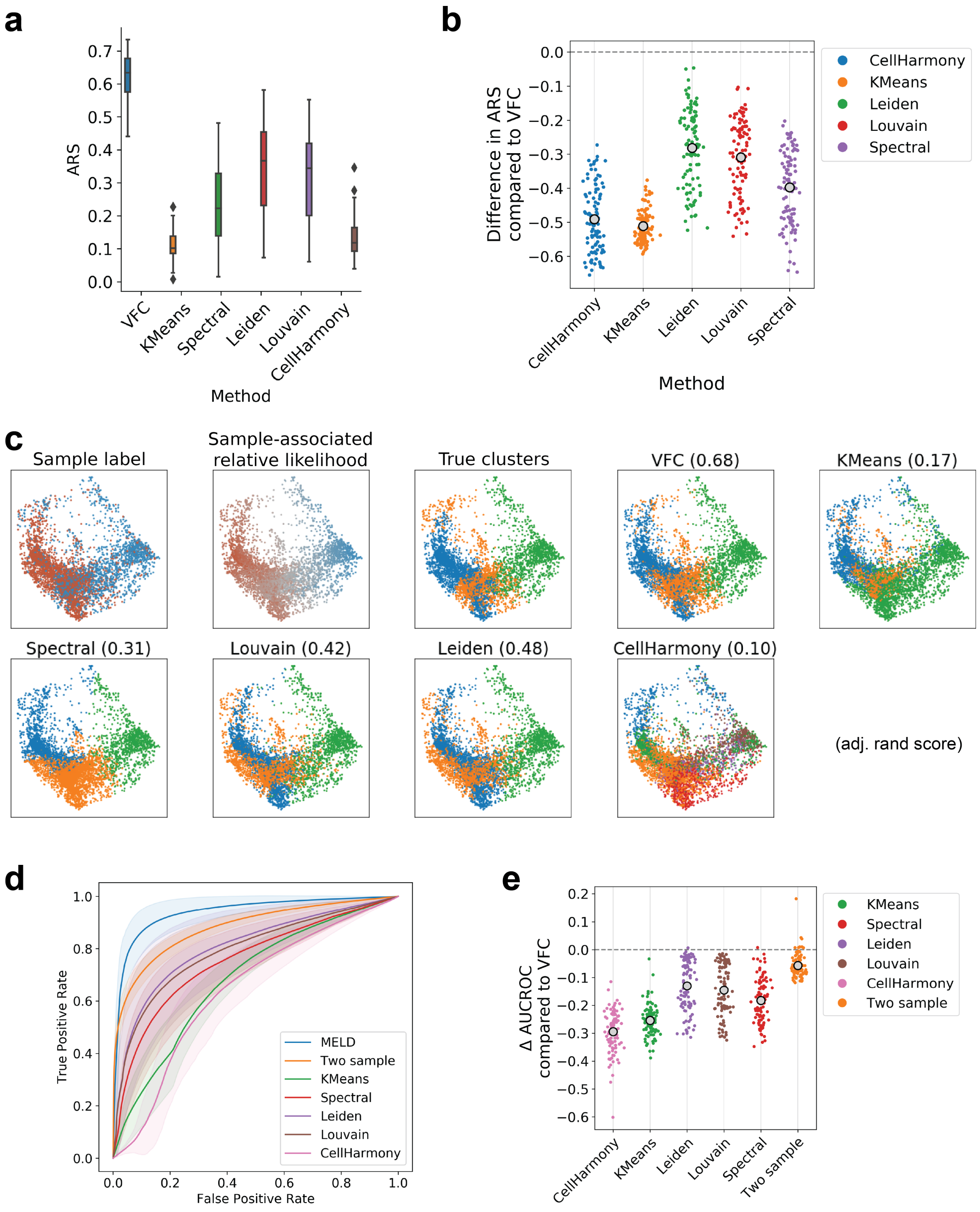


Figure S6: VFC accurately identifies cell populations affected by a perturbation in T cell data from Datlinger et al. [1]. (**a**) To create ground truth clusters, we artificially enriched and depleted various cell populations in either the experimental or control condition. Here we show the Adjusted Rand Score (ARS) over 100 simulations for 6 methods. For ARS, values close to 1 indicate perfect correspondence with ground truth, and values close to 0 indicate random labelling. VFC is the top performing method. (**b**) Because each simulation produced varying ARS scores for each method due to random seeds, we also consider the difference on performance between each method and VFC on each simulation. In none of 100 random seeds did any method outperform VFC. (**c**) The sample labels, sample-associated relative likelihoods, and clustering results for one randomly selected simulation. (**d**) Receiver operating characteristic (ROC) curves for the gene expression signatures described in the quantitative comparison section. The Area Under the Curve of the ROC (AUCROC) indicates the overall performance of each strategy for identifying a gene signature. MELD is the top performing approach followed by direct comparison of the two samples. (**e**) As above, we consider the difference in AUCROC over each of 100 simulations between MELD and each method. In only 4 simulations does another method outperform MELD by more than 0.01.

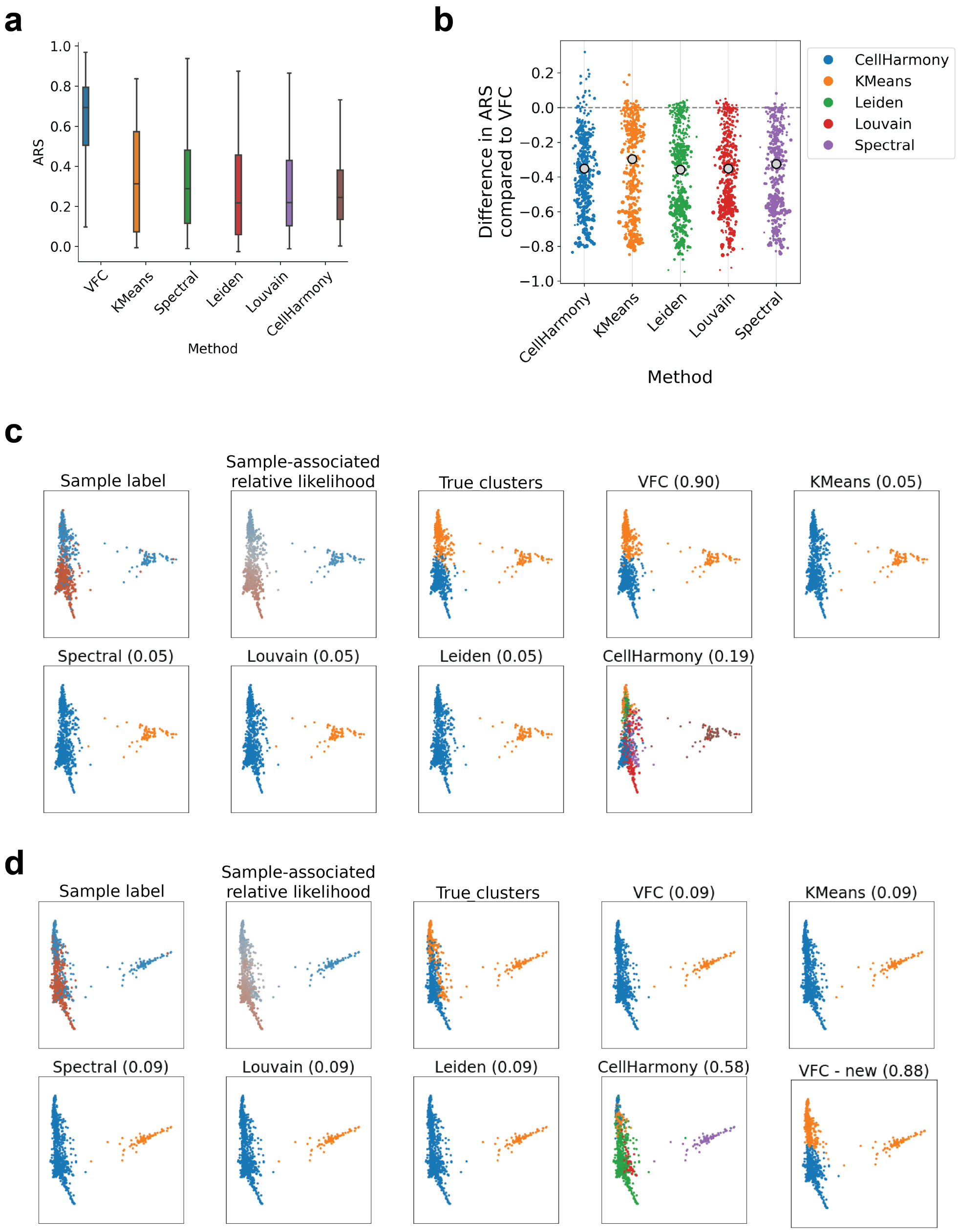


Figure S7: Quantitative comparison of clustering algorithms using zebrafish data from Wagner et al. [4]. (**a**) To create ground truth clusters, we artificially enriched and depleted various cell populations in either the experimental or control condition. Here we show the Adjusted Rand Score (ARS) over 100 simulations for 6 methods. VFC is the top performing method on average. (**b**) Difference on performance between each method and VFC on each simulation. (**c**) The sample labels, sample-associated relative likelihoods, and clustering results for the simulation in which VFC performed best relative to other methods and (**d**) for the simulation in which VFC performed worst relative to other methods. We found that by adjusting the weighting of the sample-associated relative likelihood from 1 (default) to 2, VFC becomes the top performing algorithm on this case ('VFC - new').

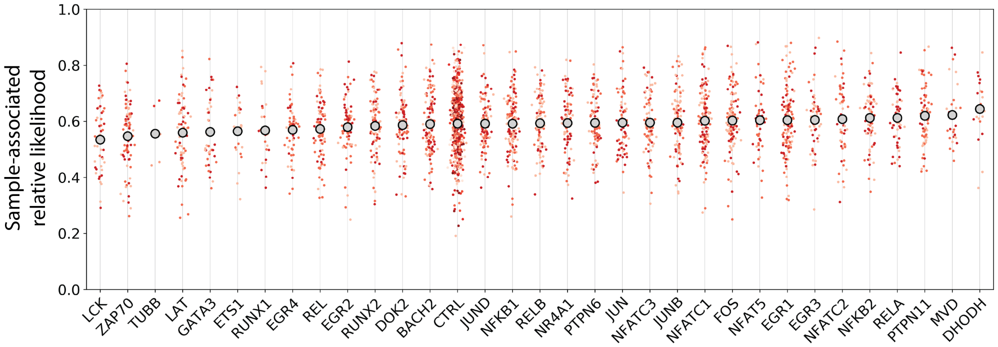


Figure S8: Quantitative analysis of Cas9 perturbations in T cells [1] using the MELD. Each plot shows the distribution of sample-associate relative likelihood values for all stimulated cells transfected with gRNAs targeting a specific gene. The shade of each cell indicates the different gRNAs targeting the same gene. To determine the impact of the gRNA on the TCR activation pathway, we rank each gene by the average stimulation likelihood value. We observed a large variation in the impact of each gene knockout consistent with the published results from Datlinger et al. [1]. Encouragingly, our results agree with their bulk RNA-seq validation experiment showing greatest depletion of TCR response with knockout of kinases LCK and ZAP70 and adaptor protein LAT. We also find a slight increase in stimulation likelihood values (and therefore stimulation) in cells in which negative regulators of TCR activation are knocked out, including PTPN6, PTPN11, and EGR3.

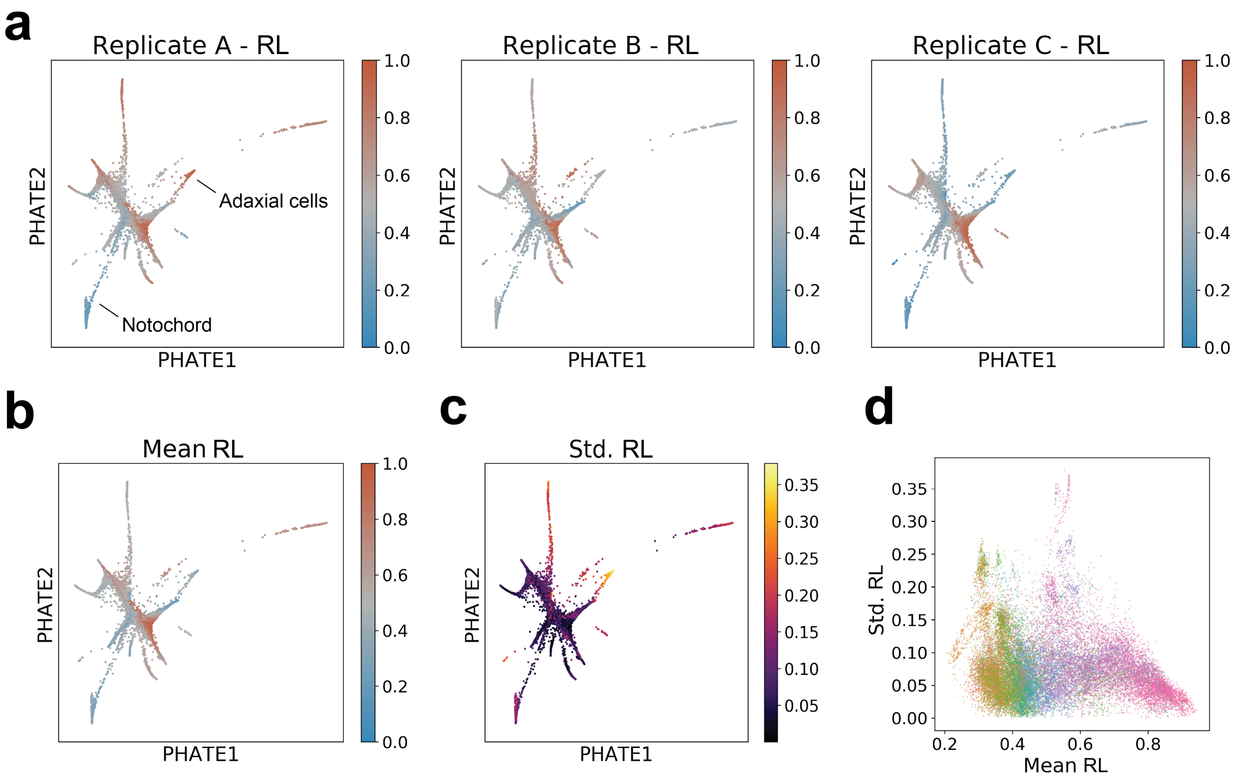


Figure S9: Analysis of replicates within the zebrafish data generated by Wagner et al. [4]. (**a**) Because the sample-associated relative likelihood (RL) is calculated by independently filtering a one-hot indicator vector for each condition, to calculate the chordin likelihood for each replicate, we simply row-normalize the smoothed vectors for the two signals indicating matched experimental / control pairs. For example, the ”Replicate A - RL” is calculated by normalizing the ”chdA” and ”tyrA” filtered indicator vectors. We notice comparing replicates that the chordin likelihood for a given cell population may vary. For example, the Adaxial cell population in enriched in the Chd condition in Replicate A, but depleted in Replicate C. Similarly, cells in the Notochord population are depleted in the Chd condition in Replicates A and C, but show minimal change in abundance in Replicate B. (**b**) The average relative likelihood across all replicates is shown for each cell on a PHATE embedding. (**c**) The standard deviation of the sample-associated relative likelihood across all replicates is shown for each cell on a PHATE embedding. Regions that have higher values exhibit greater variation in their response to the experimental perturbation. We should trust the average relative likelihood values for these cells less than for cells with little variation in relative likelihood values. (**d**) A biaxial scatter plot showing the relationship between mean and standard deviation in the relative likelihood for each cell. Color indicates the cluster labels from **Figure 5a** We observe that for cells with the highest relative likelihood, the standard deviation is smaller than for cells with relative likelihood values close to 0.5 creating a slight negative Pearson correlation of -0.18.

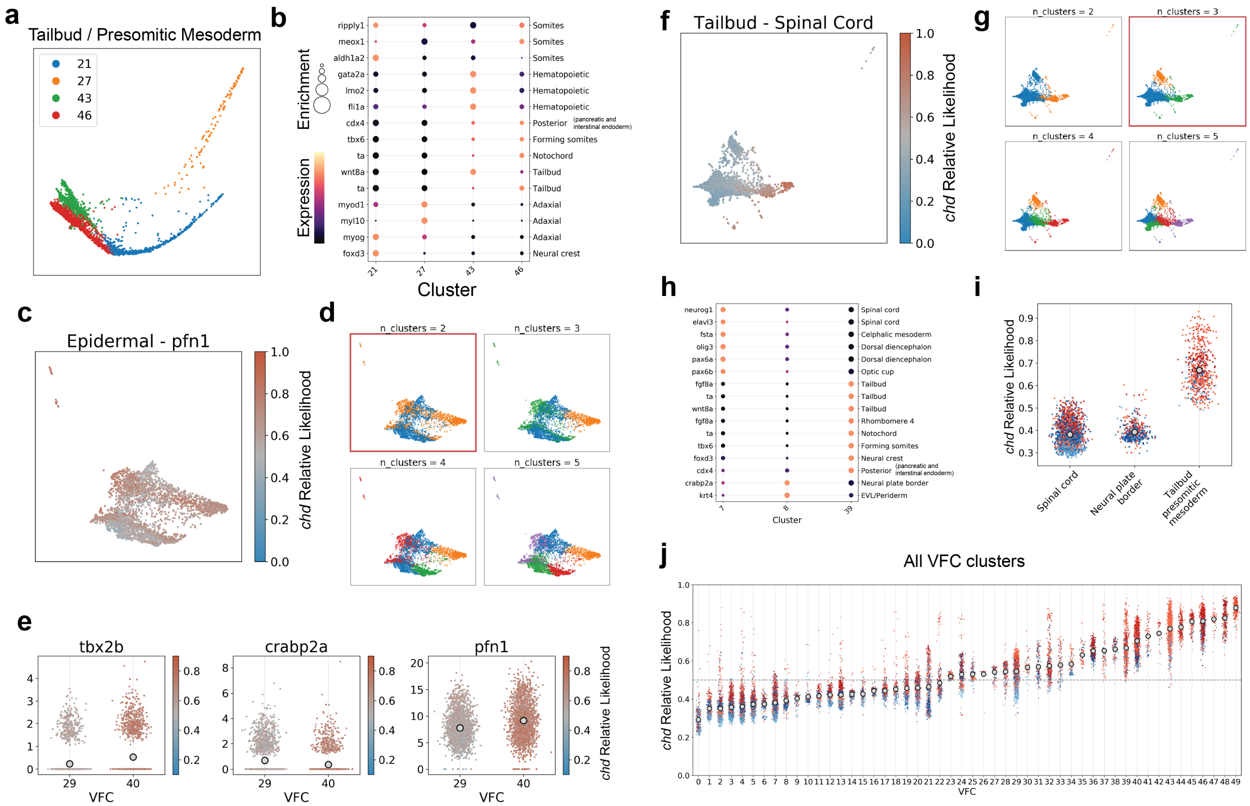


Figure S10: Characterization of vertex-frequency clusters in the zebrafish dataset. (**a**) Raw vertex-frequency cluster assignments on a PHATE visualization of the Tailbud - Presomitic Mesoderm cluster. (**b**) Normalized expression of previously identified marker genes of possible subtypes of the Tailbud - Presomitic Mesoderm [5]. The color of the dot for each gene in each cluster indicates the expression level and the size of the dot corresponds to the normalized Wasserstein distance between expression within cluster to all other clusters. (**c**) Distribution of chordin relative likelihood values within the "Epidermal - pfn1" cluster identified by Wagner et al. [4] shown on a PHATE plot. (**d**) Four different values of "n_clusters" that was used to create different VFC clusters with the "Epidermal - pfn1" cluster. We selected n_clusters = 2 because this identified a population of cells with similar chordin relative likelihood values and localization on the PHATE embedding. (**e**) Expression of three significantly differentially expressed genes between the two VFC subpopulations detected in the "Epidermal - pfn1" population. Tbx2b and Crabp2a were identified as markers of the epidermis and neural plate border respectively by Farrell et al. [5]. Because we observed differential expression of these two markers between the VFC subclusters suggests the "Epidermal - pfn1" cells identified by Wagner et al. [4] actually comprises cells originating from two distinct cell populations. (**f**) Distribution of chordin relative likelihood values within the "Tailbud - Spinal Cord" cluster identified by Wagner et al. [4] shown on a PHATE plot. (**g**) Four different values of n_clusters that was used to create different VFC clusters within the "Tailbud - Spinal Cord" cluster. We selected n_clusters = 3 because this identified populations of cells with similar likelihood values and localization on the PHATE embedding. (**h**) Same plot as in (**b**) for the subclusters of the "Tailbud - Spinal Cord". (**i**) Distribution of relative likelihood values within each VFC subcluster show that the three subclusters are biologically distinct with differing responses to the experimental perturbation. (**j**) Repeating the VFC subclustering process for all cells, we identified a total of 50 clusters within the zebrafish dataset generated by Wagner et al. [4]. Compared to the plot in **Figure 5b**, we observed a more restricted distribution of chordin relative likelihood values within each cluster suggesting these labels represent populations of cells that are more homogeneous with respect to the experimental perturbation.

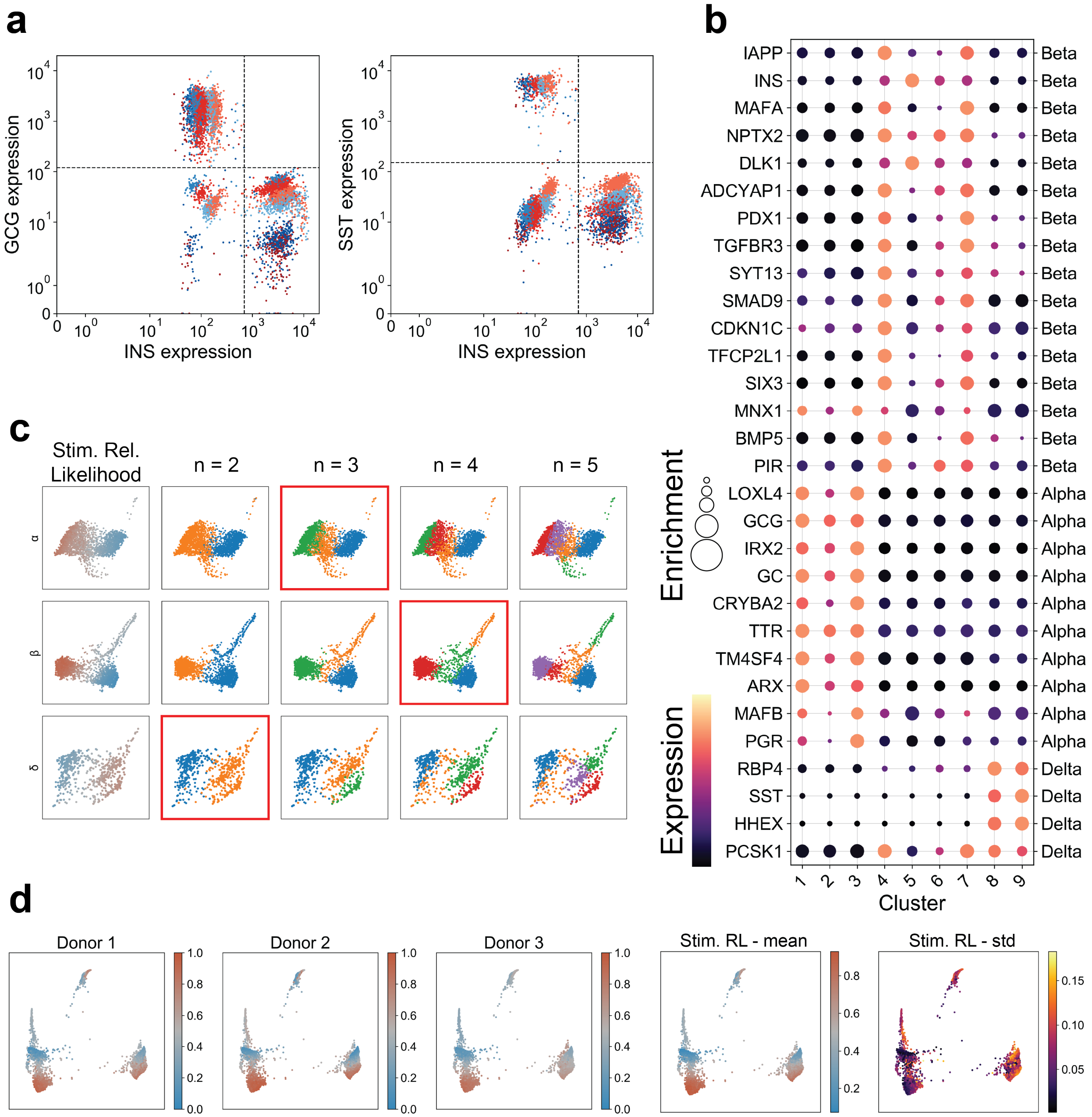


Figure S11: Analysis of pancreatic islet cells from three donors. (**a**) Library-size normalized expression of insulin (INS), glucagon (GCG), and somatostatin (SST) shows donor-specific batch effect across islet cells. (**b**) Normalized expression of previously identified marker genes of alpha, beta, and delta cells[6] 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 VFC using varying numbers of clusters for each of the three cell types. The red box denotes the selected level of clustering for each cell type. (**d**) The sample-associated relative likelihood is calculated independently for each donor and then averaged to obtain the stimulated relative likelihood used in the main analysis. We also calculate the standard deviation of the relative likelihood for each cell.

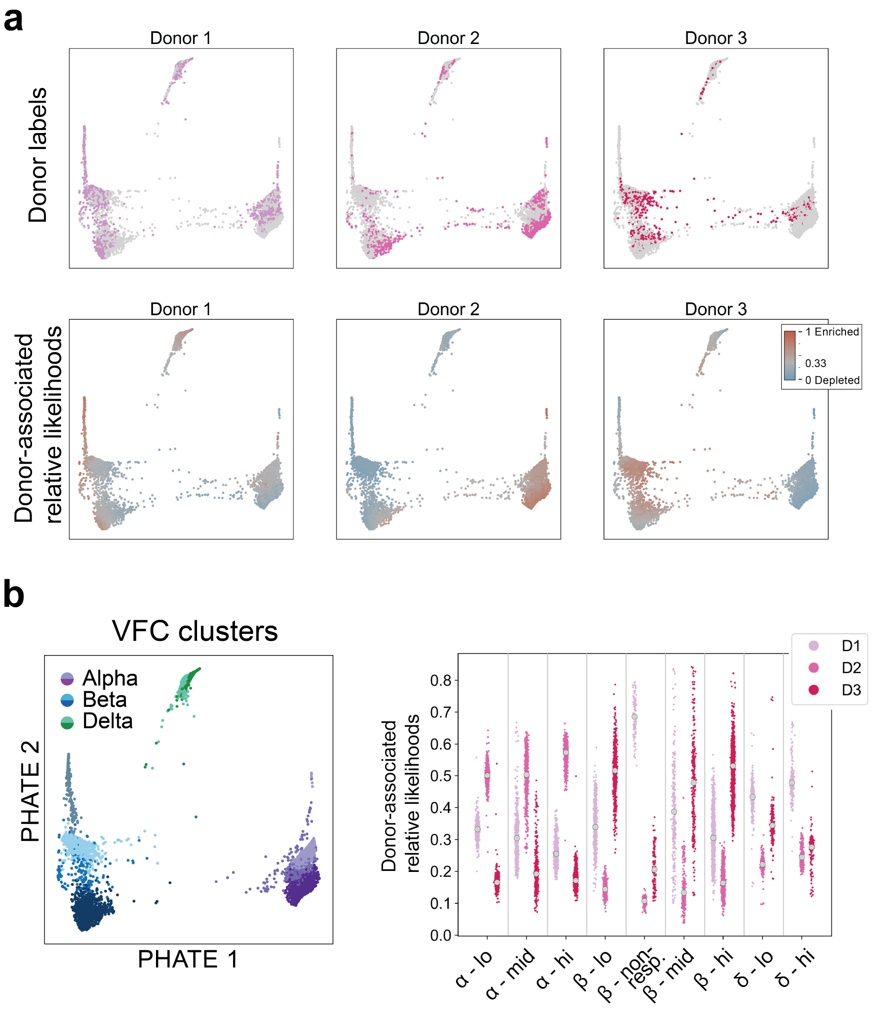


Figure S12: Analysis of islet cell profiles across donors. (**a**) The sample labels and sample-associated relative likelihood associated with each donor from which islet cells were obtained. (**b**) Comparison of the donor likelihood values within each vertex frequency cluster identifies changes in enrichment for each cluster in various donors. For example, the β - non-responsive cluster is strongly enriched in donor 1.

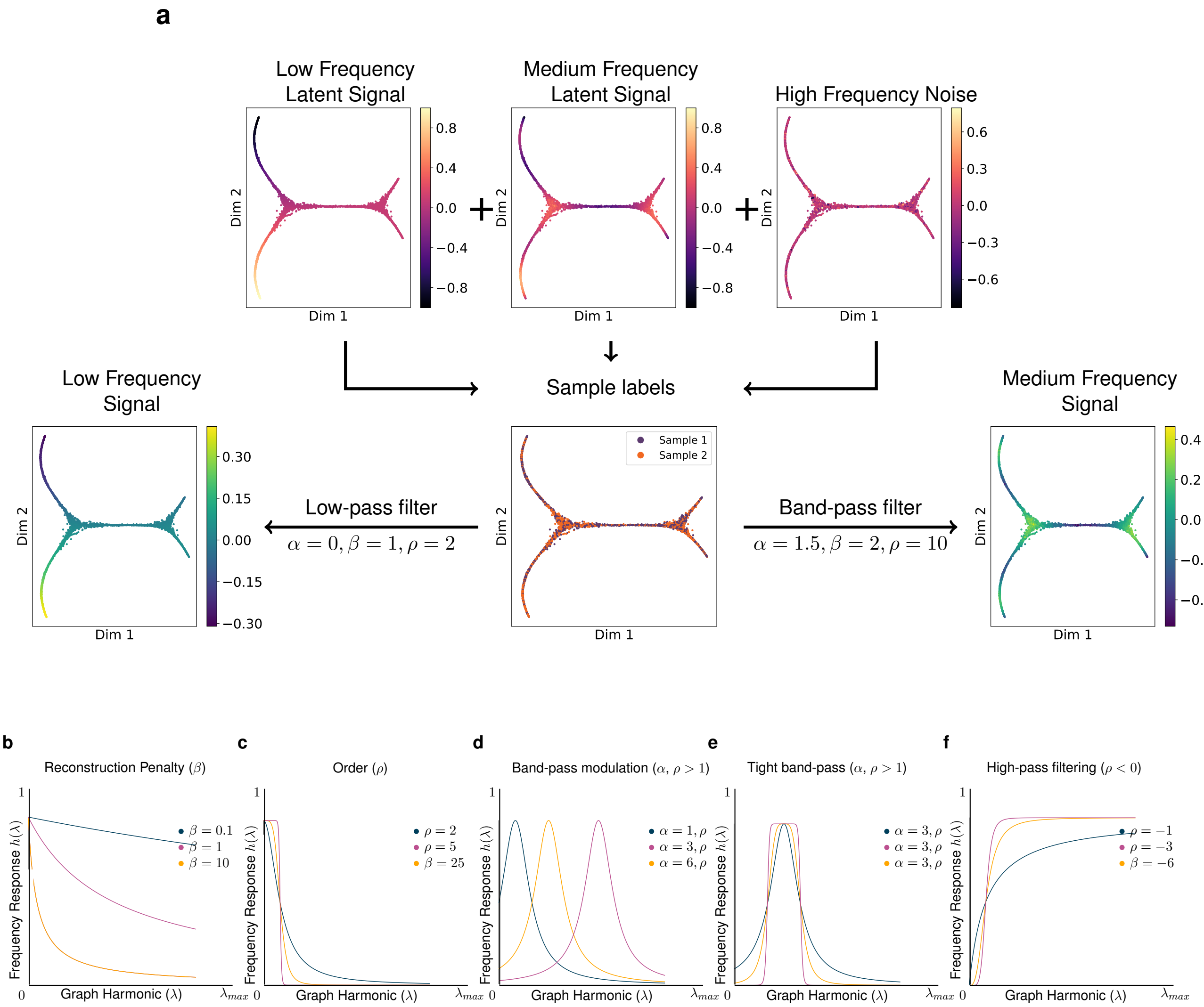


Figure S13: Source Separation and Parameter Analysis with the MELD filter. (**a**) Sample labels (center) are obtained that are a binarized observation of a low frequency latent signal (top left), a medium frequency latent signal (top middle), and high frequency noise (top right). Analysis of the sample labels alone is intractable as they are 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 and the laplacian filter is used for illustrative purposes. (**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.

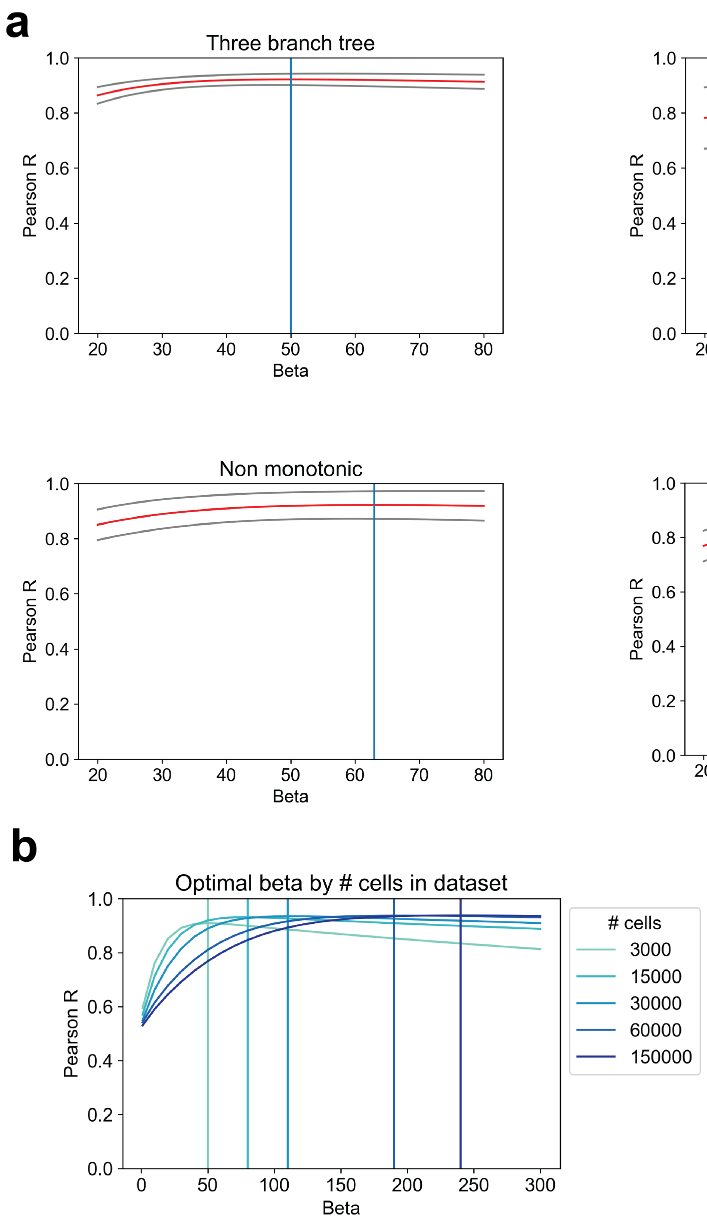


Figure S14: Selecting parameters for MELD. (**a**) Results of a parameter search over the β parameter using the four datasets described in the quantative comparisons section. The red line shows the average performance over 10 different datasets of each geometry with one standard deviation marked by the grey lines. We observe reasonably consistent performance of the sample-associated relative likelihood algorithm across all datasets using a β value between 50-75. We chose a value of 60 as the default in the MELD package and used this setting for all experiments. (**b**) We observe that the optimal β parameter for a dataset varies with the number of cells in the dataset. We suggest increasing the default beta parameter for datasets larger than 30,000 cells.

Dataset	Rel. Likelihood	Graph Averaging	kNN Averaging
Branch and Cluster	**0.82 (0.05)**	0.41 (0.05)	0.73 (0.04)
Non-monotonic	**0.94 (0.03)**	0.52 (0.06)	0.85 (0.03)
Four clusters	**0.91 (0.06)**	0.44 (0.07)	0.76 (0.07)
Three Branches	**0.90 (0.03)**	0.48 (0.07)	0.73 (0.07)
T cells [1]	**0.98 (0.01)**	0.72 (0.06)	0.32 (0.04)
Zebrafish [4]	**0.98 (0.01)**	0.53 (0.07)	0.80 (0.07)

Table S1: Quantitative comparison of methods for label smoothing over a graph. 40 random seeds were used for each of 4 synthetic datasets. 100 random seeds were used to create sample assignments on the T cell and zebrafish datasets. Average Pearson Correlation with ground truth signal is displayed with standard deviation in parentheses. Top performing algorithm is bolded.

Dataset	VFC	Spectral	Louvain	Leiden	KMeans	CellHarmony
T cell [1]	**0.62 (0.07)**	0.23 (0.11)	0.31 (0.13)	0.34 (0.14)	0.11 (0.04)	0.13 (0.05)
Zebrafish [4]	**0.53 (0.31)**	0.13 (0.15)	0.23 (0.22)	0.19 (0.21)	0.23 (0.20)	0.22 (0.16)

Table S2: Quantitative comparison of clustering methods to identify the cell types affected by a simulated experimental perturbation using real world data.

Supplementary Notes

Supplementary Note 1: A pipeline for analyzing single cell data using MELD

Using the MELD algorithm and VFC, it is now possible to propose a novel framework for analyzing single cell perturbation experiments. The goal of this framework is to identify populations of cells that are the most affected by an experimental perturbation and to characterize a gene signature of that perturbation. A schematic of the proposed pipeline is shown in **Figure S4**.

Prior to using the algorithms in MELD, we recommended first following established best practices for analysis of single cell data including exploratory analysis using visualization, preliminary clustering, and cluster annotation via differential expression analysis [2]. These steps ensure that the dataset is of high quality and comprises the cell types expected from the experimental setup. Following exploratory characterization, we propose the following analysis:

1. Estimate the sample-associate relative likelihood for each condition
2. Determine which exploratory clusters require subclustering with VFC by examining the likelihood distribution within each cluster, a visualization of the cluster, and the results of VFC with varying numbers of clusters
3. Create new cluster assignments using VFC
4. Annotate each cluster following best practices [2]
5. Characterize enrichment of cell populations using sample likelihood and gene signatures

The basic steps to calculate the sample-associated relative likelihood are provided in the Results. In the case of multiple replicates, we recommend calculating the sample density for each sample over a graph of all cells from all samples so long as there is sufficient overlap between samples. This overlap can be assessed using the k-nearest neighbor batch effect test described in Büttner et al. [7]. We then normalize the sample density for matched experimental and control samples of the same replicate and average across replicates to obtain an average measure of the perturbation. Variation in this likelihood across replicates can be used as a measure of consistency for the measured perturbation across cell types. The result of this step is an estimate of the probability that each cell would be observed in the treatment condition relative to the control.

Having calculated the sample-associated relative likelihood, we next recommend determining which cell populations identified during exploratory analysis require further subclustering with VFC to identify cell types enriched or depleted in the experimental condition. Determining optimal cluster resolution for single cell analysis will vary across experiments depending on the biological system being studied and the goals of each individual researcher. Instead of providing a single measure to determine the number of clusters, we outline a general strategy as a guide for users of MELD.

To determine the number of VFC clusters, we suggest taking into consideration transcriptional variation within each coarse-grained cluster and the effect of the perturbation. First, using a dimensionality reduction tool such as PHATE, examine a two or three dimensional scatter plot of the cluster colored by the sample likelihood for each cell. Here, the goal is to identify either regions that have very different likelihood values or regions of data density separated by low-density regions suggesting the present of multiple subclusters to target with VFC. We also suggest examining the distribution the likelihood values within each cluster to determine if the cells in the cluster exhibit a restricted range of responses to the perturbation or large variation that would require subclustering. Finally, we recommend running VFC with various numbers of clusters (2-5 is often sufficient) and inspecting the output on a PHATE plot and/or with a swarm plot. In ambiguous cases, it may be helpful to perform differential expression analysis and gene set enrichment to determine whether or not each cluster is biologically relevant to the experimental question under consideration [2, 3]. Importantly, not all clusters need subclustering, and we emphasize the ideal cluster resolution will vary based on the goals of each analyst.

To determine the gene signature of the perturbation, we recommend quantifying the differences in expression between VFC clusters. For experiments with only a single cell type and 3-4 VFC clusters, it is often sufficient to perform differential expression analysis between the cluster most enriched in the experimental condition and the cluster most depleted in the experimental condition. And example of this analysis is provided in the T cell analysis section of the **Results**. For experiments with several cell types, we recommend calculating the gene signature between the enriched and depleted VFC clusters within each exploratory cluster. To obtain a consensus gene signature, a research may take the intersection of the gene signatures within exploratory cluster. An example of this analysis is provided in the pancreatic islets section of the **Results**.

We note that the strategy for identifying gene signatures outlined in the previous paragraph differs from the current framework employed in recent papers (**Figure S3**). Instead of comparing expression between cells from the experimental condition and the control, we compare clusters of cells identified with VFC. The rationale for the framework presented here is that if VFC clusters are transcriptionally homogeneous and exhibit a uniform response to the perturbation, we expect differences in gene expression between conditions *within* each cluster to represent biological and technical noise. However, characterizing transcriptional differences *between* cells of different clusters regardless of condition of origin will yield a description of the cell states that vary between experimental conditions. We confirm that the gene signatures obtained in this manner are more accurate than between-sample comparisons in our quantitative comparisons.

Supplementary Note 2: VFC improves analysis of *chd* Cas9 knockout in zebrafish embryos Here we provide details of our analysis of three clusters in the zebrafish datasets [4] that required further subclustering using VFC. In each example, we show biologically relevant insights that were missed in the published analysis.

The Tailbud – Presomitic Mesoderm (TPM) cluster exhibits the largest range of chordin relative likelihood values of all the clusters annotated by Wagner et al. [4]. In a PHATE visualization of the cluster, we observe many different branches of cell states, each with varying ranges of chordin relative likelihood values (**Figure 5c**). Within the TPM cluster, we find four subclusters using VFC (**Figure 5d**). Using established markers [5], we identify these clusters as immature adaxial cells, mature adaxial cells, presomitic mesoderm cells, and hematopoietic cells (**Figures 5c & S10**). Examining the distribution of chordin relative likelihood scores within each cell type, we conclude that the large range of chordin relative likelihood values within the TPM cluster is due to largely non-overlapping distributions of scores within each of these subpopulations (**Figure 5e**). The immature and mature adaxial cells, which are embryonic muscle precursors, have low chordin relative likelihood values indicating depletion of these cells in the *chd* condition which matches observed depletion of myotomal cells in chordin mutants [8]. Conversely, the presomitic mesoderm and hematopoietic mesoderm have high chordin relative likelihood values, indicating that these cells are enriched in a chordin mutant. Indeed, expansion of the hematopoietic mesoderm has been observed in chordin morphants [9] and expansion of the presomitic mesoderm was observed in siblings of the *chd* embryos by Wagner et al. [4]. This heterogeneous effect was entirely missed by the fold-change analysis, since the averaging of all cells assigned to the TPM cluster caused the depletion of adaxial cells to be masked by the expansion of the presomitic and hematopoietic mesoderm.

Another advantage of vertex-frequency clustering is that we can now differentiate between a change in gene expression levels across conditions and a change in abundance of cells expressing a given gene between conditions. When we examined marker gene expression within each of the VFC subclusters, we find different trends in expression in each cluster (**Figure 5f**). For example, Myod1, a marker of adaxial cells, is lowly expressed in the presomitic and hematopoietic mesoderm, but highly expressed in adaxial cells. Using a rank sum test, we find that Myod1 is not differentially expressed between conditions within any of the VFC clusters despite there being differential expression using all cells in the TPM cluster (**Figure 5f**). We find a similar trend with Tbx6, a mesoderm marker that is not expressed in adaxial cells. We find Tbx6 is differentially expressed between *chd* and *tyr* embryos within the whole cluster but not within the adaxial or presomitic mesoderm clusters. These results show that the observed change in expression of these genes in the published analysis was in fact due to changes in abundance of cell subpopulations that led to misleading differences in statistics calculated across multiple populations as a whole. Using the chordin relative likelihood and VFC, we can identify more appropriate clusters.

We similarly analyzed the ”Epidermal - pfn1 (EPP)” and ”Tailbud - Spinal Cord (TSC)” clusters which had the 6th and 3rd largest standard deviation in chordin relative likelihood values of all published clusters, respectively (**Figure S10**). We used VFC to break up the Epidermal - pfn1 cluster into two subclusters. Among the top differentially expressed genes between the resulting clusters we find tbx2b, crabp2a, and pfn1. Crabp2a, a marker of the neural plate border [5], is more lowly expressed in the cluster with higher chordin relative likelihood values, suggesting that *chd* loss-of-function inhibits expression of crabp2a. This is consistent with previous studies showing a requirement of chordin for proper gene expression patterning within the neural plate [10, 11].

Within the Tailbud - Spinal Cord cluster we further identified three subpopulations of cells using VFC. Examining gene expression within the subclusters, we can see that the published cluster contains different populations of cells. One group expresses markers of the spinal cord (neurog, elavl3) and dorsal tissues (olig3, pax6a/b) with an average chordin relative likelihood of 0.38, which is consistent with prior evidence that *chd* loss-of-function disrupts specification of the neuroectoderm and dorsal tissues such as the spinal cord [8]. Examining the two remaining subclusters, we see that these cells resemble cells found in both the TPM and Epidermal - Pfn1 clusters. One cluster exhibits high levels of crabp2a and chordin relative likelihood values <0.5 similar to the neural plate border cells subpopulation within the Epidermal - Pfn1 cluster. Similarly, we find the remaining cluster expressed markers of the tailbud and presomitic mesoderm including tbx6, sox2, and fgf8a. Together, these results demonstrate the advantage of using the sample-associated relative likelihood and vertex frequency clustering to quantify the effect of genetic loss-of-function perturbations in a complex system with many cell types.

Supplementary Note 3: Applying MELD analysis to single cell datasets with a batch effect

When jointly analyzing single cell datasets collected in different samples, difficulty may arise due to systematic changes in gene expression profiles between biologically equivalent cells [7]. These changes may be technical in nature (e.g. differences in the reverse transcription efficiency during library preparation) or biological (e.g. changes in sample preparation cause unexpected changes in biological state of otherwise equivalent cells). Regardless of the cause, the unifying feature of batch effects is that they confound the analysis a given research wants to perform. As such, it is unsurprising that dozens of batch normalization tools have been developed for single cell data [12]. However, it is important to emphasize that what constitutes a batch effect is dependent on the biological question in which a researcher is interested. Some analysts might be uninterested in variation caused by a change in cell media composition between samples, but other researchers might want to study these differences. Batch normalization tools have no way to know what variation is biologically relevant to the specific hypotheses of a given experiment and thus risk removing meaningful experimental signal when ”correcting” measured values. This is problematic for analysis using MELD, because the goal of the toolkit is to quantify the differences that exist between samples without regard for the specific interests of given hypothesis. As such, we do not recommend using batch correction along the experimental axis (i.e. between experimental and control conditions) before running MELD. However, recognizing that in some cases batch correction is essential, we describe several considerations for performing MELD analysis on batch-corrected data.

For the MELD algorithm to accurately estimate relative likelihood for each sample, we assume that the graph learned from single cell data approximates the underlying cell state manifold. In the **Methods** we describe the use of an anisotropic kernel that normalizes for varying sampling density across cell states. However, some batch correction methods, such as mutual nearest neighbors [13], rely on the construction of a graph with artificially inflated weights between nodes from different samples. This graph no longer models the cell states an experiment measured, but rather enforces similarities between cells based on the heuristic of the chosen normalization model. We provide no theoretical guarantees that a graph learned from batch corrected data will accurately model the underlying probability densities of each condition.

In practice when analyzing islet cells collected from multiple donors, that applying batch correction methods across the donor label improves our ability to capture a signal of IFNg stimulation. It is important to note that in this case, batch correction applied to a label that is orthogonal to the experimental axis. We have no examined the accuracy of the MELD algorithm when batch correction is applied between experimental and control samples, although it is our expectation that this will likely remove biological signal. We recommend any user considering applying batch correction methods prior to running MELD analysis follow these steps:

1. To determine if a batch effect exists, confirm that cells from one sample are not finding appropriate neighbors in another following the strategy outlined by Büttner et al. [7].
2. To characterize the effect, identify which genes change the most between the samples
3. Confirm that the genes that are different are not relevant to the biological question under investigation
4. Apply batch correction
5. Confirm that relevant biological differences are still present using MELD analysis
6. If the biological differences are not present, repeat from step 1 with less batch correction. If you hit your personal recursion limit, consider that you don't actually want to do batch correction
7. If biological differences are present, then confirm that previous batch effect has been corrected and proceed to downstream analysis

References

[1] 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.

[2] 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.

[3] Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L. Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo, and Stephanie C. Hicks. Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, 17(2):137–145, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0654-x.

[4] 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.

[5] 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.

[6] 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.

[7] Maren Büttner, Zhichao Miao, F. Alexander Wolf, Sarah A. Teichmann, and Fabian J. Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16(1):43–49, January 2019. ISSN 1548-7105. doi: 10.1038/s41592-018-0254-1.

[8] 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.

[9] Anskar Y. H. Leung, Eric M. Mendenhall, Tommy T. F. Kwan, Raymond Liang, Craig Eckfeldt, Eleanor Chen, Matthias Hammerschmidt, Suzanne Grindley, Stephen C. Ekker, and Catherine M. Verfaillie. Characterization of expanded intermediate cell mass in zebrafish chordin morphant embryos. *Developmental Biology*, 277(1):235–254, January 2005. ISSN 0012-1606. doi: 10.1016/j.ydbio.2004.09.032.

[10] Ben Steventon, Claudio Araya, Claudia Linker, Sei Kuriyama, and Roberto Mayor. Differential requirements of BMP and Wnt signalling during gastrulation and neurulation define two steps in neural crest induction. *Development*, 136(5):771–779, March 2009. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.029017.

[11] Carolin Schille and Alexandra Schambony. Signaling pathways and tissue interactions in neural plate border formation. *Neurogenesis*, 4(1), February 2017. ISSN 2326-2133. doi: 10.1080/23262133.2017.1292783.

[12] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, page 2020.05.22.111161, May 2020. doi: 10.1101/2020.05.22.111161.

[13] 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.