Causal identification of single-cell experimental perturbation effects with CINEMA-OT

Mingze Dong[1,2], **Bao Wang**[3,4], **Jessica Wei**[4,5], **Alexander Frey**[4,5], **Feriel Ouerghi**[4,5], **Ellen F. Foxman**[3,4†], **Jeffrey J. Ishizuka**[2,4,5†], **Rahul M. Dhodapkar**[6†], **David van Dijk**[1,7,8†]
[1] Interdepartmental Program in Computational Biology & Bioinformatics, Yale University
[2] Department of Pathology, Yale School of Medicine
[3] Department of Laboratory Medicine, Yale School of Medicine
[4] Department of Immunobiology, Yale School of Medicine
[5] Department of Medical Oncology, Yale School of Medicine
[6] Department of Ophthalmology and Visual Science, Yale School of Medicine
[7] Department of Internal Medicine (Cardiology), Yale School of Medicine
[8] Department of Computer Science, Yale University
† Co-corresponding authors
Correspondence should be addressed to: david.vandijk@yale.edu

Abstract

Recent advancements in single-cell technologies allow characterization of experimental perturbations at single-cell resolution. While methods have been developed to analyze data from such experiments, the application of a strict causal framework has not yet been explored for the inference of treatment effects at the single-cell level. In this work, we present a causal inference based approach to single-cell perturbation analysis, termed CINEMA-OT (Causal INdependent Effect Module Attribution + Optimal Transport). CINEMA-OT separates confounding sources of variation from perturbation effects to obtain an optimal transport matching that reflects counterfactual cell pairs. These cell pairs represent causal perturbation responses permitting a number of novel analyses, such as individual treatment effect analysis, response clustering, attribution analysis, and synergy analysis. We benchmark CINEMA-OT on an array of treatment effect estimation tasks for several simulated and real datasets and show that it outperforms other single-cell perturbation analysis methods. Finally, we perform CINEMA-OT analysis of two newly-generated datasets: (1) rhinovirus-infected airway organoids, and (2) combinatorial cytokine stimulation of immune cells. Using CINEMA-OT, we discover diverging treatment responses and their associated cellular subpopulations. By applying CINEMA-OT to combinatorial experimental designs, we infer the specific cell-gene programs driving syngergistic responses.

1 Introduction

Cellular responses to environmental signals are a fundamental component of biological functioning, playing an integral role in both homeostasis and disease [1]. For decades, controlled perturbation experiments have been used to reveal the underlying mechanisms of biological processes. Recent advances in single-cell technologies allow complex experiments measuring high dimensional phenotypes at high throughput under diverse stimulation conditions [2–8]. However, deriving biological insights from these experiments remains a challenge.

For the analysis of single-cell perturbation data, several approaches have been developed, which can be categorized as follows:

1. *Differential expression* methods test for statistically significant differences in gene expression between cell populations. [9–13]

2. *Differential abundance* methods formulate the quantification problem as differential abundance testing in a continuous cellular state manifold. [14–17]

3. *Perturbation analysis* methods aim to reveal underlying biological processes by modeling perturbation in a latent space with either linear models or neural networks. [3–7, 18–21]

While techniques to characterize the effects of perturbations by averaging over populations have been extensively used in the analysis of single-cell data, methods allowing for causal single-cell perturbation analysis are lacking. Causal inference is a field of active research aiming to solve the general problem of response quantification [22]. In causal inference, perturbations are considered *treatments* and the general problem of modeling response to perturbation is known as the *treatment effect estimation problem*. In subsequent portions of this manuscript, we will borrow from the terminology of causal inference, referring to perturbations and treatments, as well as response and treatment effect, interchangeably. Ideal causal methods allow for the direct characterization of underlying confounding variation, which is not accounted for in any of the existing single-cell analysis methods.

We consider confounding variation to be causal of differential response, as well as differential underlying phenotype [23]. In the case of scRNA-seq experiments, sources of variation such cell cycle stage, microenvironment, and pre-treatment chromatin accessibility may all act as confounding factors when performing treatment effect estimation [18]. Collectively, confounding factors can be thought of as a cell's context that may both influence a cell's underlying gene expression profile, and condition treatment-induced gene signatures. If confounders are incorrectly identified, counterfactual cell pairs will be inappropriately matched, and treatment effects cannot be correctly estimated.

One well-established confounding factor that may affect treatment response is cell type. For example, widely used nucleoside analog chemotherapeutics such as 5-fluorouracil (5-FU) act selectively on cells in the DNA synthesis phase of the cell cycle, killing cancer cells while minimizing effects on healthy tissue [24]. Some mutations may also drive differential response to a stimulation, as is seen with some tumors in response to TGF-β [25]. Confounders may be latent, such as different exposures of cells to a drug, which may have different effects at different concentrations within each individual cell. Thus, our framework must be able to account for all of these types of confounders.

Our solution is to introduce a causal framework permitting explorations of perturbation effects and how they may richly interact with confounder states. By explicitly modeling the diversity of cellular responses to perturbations across confounding factors without a requirement to pre-identify cell populations of interest, we may identify cell state-conditioned transcriptional changes driving pathology or pharmacologic response to perturbation.

In this paper, we present CINEMA-OT (Causal INdependent Effect Module Attribution + Optimal Transport), which applies independent component analysis (ICA) and filtering based on a functional dependence statistic to identify confounding factors and treatment-associated factors. CINEMA-OT then applies weighted optimal transport (OT) to achieve causal individual matching. The algorithm is based on a causal inference framework for modeling confounding signals and conditional perturbation effects at the single-cell level, relying on two key assumptions. We show that the model is uniquely indentifiable via the theory of ICA. The computed causal cell matching enables a multitude of novel downstream analyses, including but not limited to: individual treatment effect estimation, treatment synergy analysis, sub-cluster level biological process enrichment analysis, and attribution of perturbation effects.

We demonstrate the power of CINEMA-OT by benchmarking it on several simulated and real datasets and comparing it to widely used single-cell level perturbation analysis methods. We then perform CINEMA-OT analyses of two newly-generated datasets. In the first, we examined the effects of virus infection and cigarette smoke on innate immune responses in airway organoids. In the second, we performed combinatorial cytokine stimulation of ex vivo peripheral blood mononuclear cells in order to characterize how cytokines act in concert to shape immune responses.

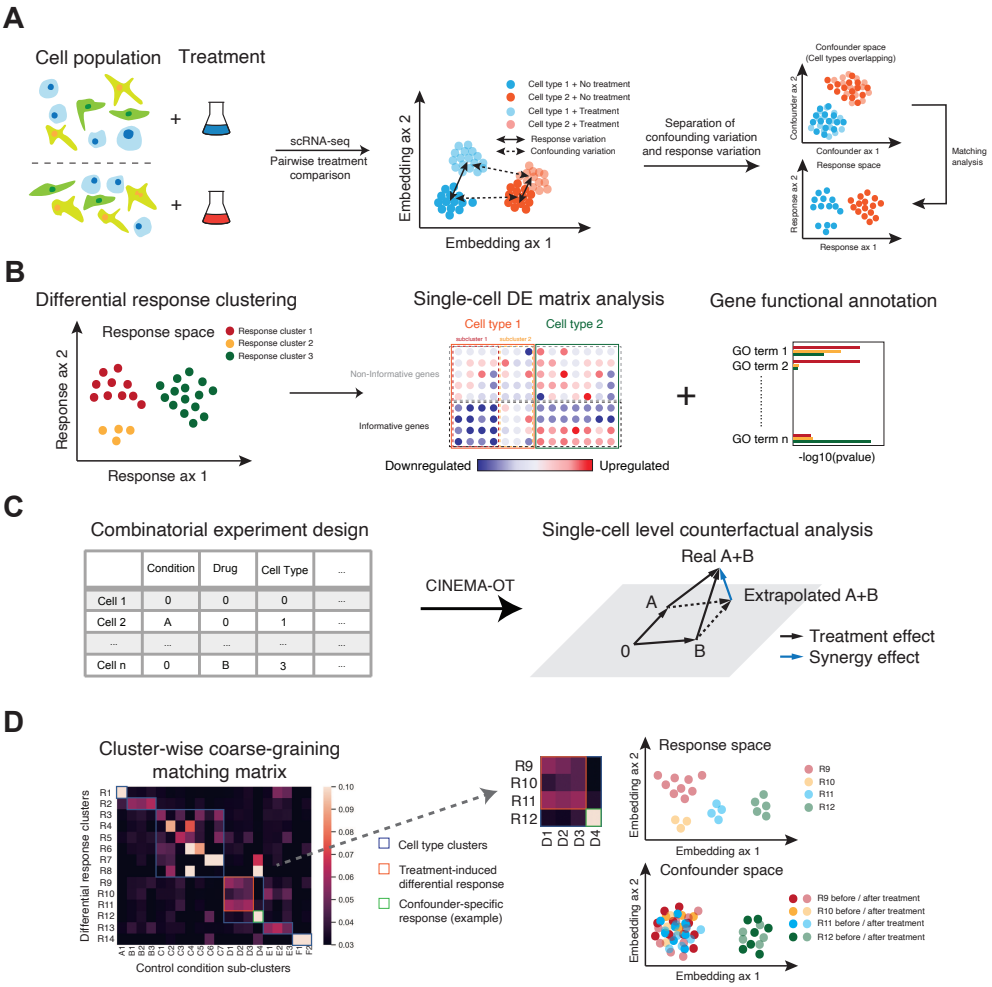

Figure 1: A causal framework for single-cell level perturbation effect analysis. **A.** In single-cell experiments, cells are separated by both treatment batches and latent cell states. Latent states of cells have confounding effects on the effects of perturbation. Upon successfully separating confounding variation and response variation in data, we can identify confounding signals for each cells and draw counterfactual pairs across cells to compute causal perturbation effects. **B.** After characterizing the differential response matrices at a single-cell level, we can subcluster cells by treatment responses. These responses may be further characterized by other tools, such as gene set enrichment analysis. **C.** We are able to quantify the synergy effect in combinatorial perturbations by evaluating the dissimilarity of extrapolated phenotypes and true combinatorially perturbed phenotypes. **D.** CINEMA-OT can attribute divergent treatment effects to either explicit confounders, or latent confounders by cluster-wise coarse-graining of the matching matrix.

2 Results

2.1 Confounder signal matching via CINEMA-OT

To perform causal single-cell perturbation effect inference, we have adopted the *potential outcome causal framework* [22, 26]. Ideally, to generate causal assertions about the effect of a perturbation on the transcriptional state of a given cell, we would like to measure the same cell both before and after the perturbation is applied. However, the process of obtaining transcript measurements from single cells is destructive, and an individual cell may only be measured once. In fact, to the best of our knowledge, all high throughput genomic measurement technologies are destructive. A solution would be to infer *counterfactual* cell pairs, which are inferred, causally-linked pairs—predictions of what a cell in one condition would look like in another condition. The potential outcome framework formalizes this concept by establishing a statistical model that describes outcome variables as a

function of confounding factors and treatment-associated factors. In our setting, to infer single-cell treatment effects, this would translate to distinguishing the effects of biological variation and treatment on treatment-associated genes.

In the potential outcome framework, the key difficulty for general unsupervised causal inference is the mixing of confounders with outcomes. In the field of causal inference, such a case is described as learning with both interventions and latent confounding, which remains an active area of research [27]. In our case, a gene can contribute to confounding variation as well as treatment-associated variation. To apply the tools of classical causal inference, confounding factors must first be distinguished from treatment-associated factors. Notably, confounding factors and treatment-associated factors may be treated as a low dimensional function of the gene space.

To unmix confounding effects and treatment-associated effects, we propose two sufficient assumptions, which can be potentially relaxed:

Assumption 1: (Independent sources and noise). *Confounding factors and treatment events are pairwise independent random variables.*

This assumption relies on treated and untreated cells being drawn from the same underlying set of cells. In practice, this is a central part of most single-cell experimental designs.

Assumption 2: (Linearity of source signal combinations). *Confounding gene signatures can be modeled as a linear combination of confounding sources plus an independent noise term. The outcome gene signatures can be modeled as arbitrary functions of confounding factors and treatment events plus an independent noise term.*

While modeling total expression as a linear combination of relevant factors may be unable to capture complex nonlinear relationships between confounding variables, this assumption is necessary in our case, as the nonlinear source separation problem is generally not identifiable. Moreover, we use the confounders as a intermediate metric to construct a matching across similar cells, therefore full characterization of nonlinear dependency is not necessary in our method. Meanwhile, our assumption allows nonlinear interactions between confounders and treatments, thereby permitting modeling of general non-linear treatment effects, including confounder-specific treatment effects.

Given these assumptions, we can strictly prove that the identification of confounding factors is equivalent to solving the blind source separation problem (BSS), which can be done by the ICA algorithm (see Methods for the proof).

After ICA transformation, The gene count matrix is decomposed into a linear combination of confounding signals, treatment-associated signals, and independent noise signals. In practice, we filter the noise signal by PCA dimensionality reduction prior to ICA. To identify treatment signals, we use a non-parametric distribution-free test based on Chatterjee's coefficient for functional dependence between each identified signal and the ground truth treatment signal [28] (Figure 2A).

Finally, with the identified confounding factors, we are able to apply a causal matching procedure, which aims to match cells according to their coordinates in the embedded confounder space. In order to match cells, k-nearest neighbor (knn) matching may be first considered because of its wide use in scRNA-seq analysis. However, in practice, mutual knn matching and other local matching techniques such as ϵ-NN may collapse matches at the boundaries of separated cell clusters, reducing their robustness to outliers (Figure 2B). By contrast, optimal transport is a mass-preserving matching procedure that does not suffer from this drawback.

While solving the optimal transport problem is often prohibitively resource-intensive for large-scale biological data, CINEMA-OT considers the tractable case of entropic regularization [29, 30]. Optimal transport with entropic regularization can be formulated as a convex optimization problem which can be solved efficiently using the alternating direction method (Sinkhorn-Knopp algorithm). Moreover, recent works have shown asymptotic properties of the entropy regularized optimal transport map for causal matching [31]. Entropy regularized optimal transport yields a probabilistic one-to-many matching for each point between two discrete distributions. The mapping generated in this way is smooth, and robust to outliers [30]. In CINEMA-OT, to achieve a mapping between cells

across treatment conditions that is not distorted by treatment effects, we perform optimal transport in the confounder space.

2.2 Differential abundance correction via causal reweighting

A treatment may change the distribution of cell densities, e.g. cells may die or proliferate in response to some perturbation. Thus, we may have an additional factor of differential confounder abundance across experimentally perturbed datasets. This factor can come from either real biological effects or batch effects. The differential abundance confounder can affect the performance of CINEMA-OT since in this case the underlying confounders are no longer independent of the treatment event and assumption (1) is violated. Indeed, our experiments have shown that while CINEMA-OT can tolerate moderate levels of differential abundance (Extended Figure 2), it can fail when high levels of differential abundance are present (Extended Figure 3).

In order to correct for the effects caused by differential abundance, we have implemented an iterative variant of CINEMA-OT. In each iteration step, we estimate overlap weights using likelihood estimation of the treatment events (for details, see Methods) (Figure 2C). By balancing the local distribution of cells in the confounder space using these weights, we can correctly match cells.

We note that this weighted version of CINEMA-OT should be used only when required. As the weighted version of CINEMA-OT relaxes assumption 1 in our framework, the identifiability of our model can no longer be guaranteed, which may reduce its ability to identify certain classes of cellular responses. For example, aligning cell types with differential abundance across treatment conditions eliminates the possibility that one cell type can convert to another cell type upon treatment. In such a case, cell type can be a treatment-induced factor instead of a confounder. Accounting for this, in order to integrate prior biological knowledge, CINEMA-OT also provides an option to assign weights according to user-provided labels (e.g. cell-types). In this case, CINEMA-OT can assign weights using confounder labels instead of automatically balancing over all possible covariates.

2.3 Analysis following CINEMA-OT

With the matched counterfactual cell pairs computed by CINEMA-OT, we are able to obtain two key outputs: (1) the matching correspondence matrix across treatment conditions, and (2) the individual treatment effect (ITE) for each cell with its counterfactual pair across treatments (Figure 1A).

Individual treatment effect matrices are gene by cell matrices which can be clustered and visualized by existing scRNA-seq computational pipelines. By clustering over an ITE matrix, we can identify sub-clusters within cell types with heterogeneous response to treatments. We may perform statistical analysis to identify the significantly affected genes and identify their coordinated biological function by gene set enrichment analysis (Figure 1B).

In addition, in a combinatorial perturbation experiment, we are able to define a synergy effect metric by comparing the predicted effect of combining multiple treatments with the observed effect of combined treatment (Figure 1C). We define a synergy metric by estimating the difference between the true sample under combinatorial treatment (A+B) and the predicted sample, by adding the effects of treatment A and treatment B, thus assuming purely linear, non-interactive effects. If no difference is measured, we may conclude that there are no nonlinear or interaction effects between the treatments. If non-zero synergy is present, this points to some interaction between treatments A and B. Synergy is computed for every cell-gene pair, resulting in a matrix of equivalent form to expression or ITE matrices - a unique feature of CINEMA-OT.

Another important task in perturbation effect analysis is attribution of treatment effects. Differential response can be driven by either differences in explicit confounding factors or latent factors, such as treatment heterogeneity. Because CINEMA-OT provides a single-cell level matching as one output, the task can be solved by coarse-graining the matching matrix. Responses that cluster both in response as well as confounder space may be attributed to explicit confounding factors. Conversely, responses that cluster well in the response space but do not demonstrate clustering in the confounder space may be attributed to latent factors (Figure 1D). Such an analysis can be performed either at the cell type level or at the sub-cluster level to reveal underlying heterogeneity. (see Methods for additional details)

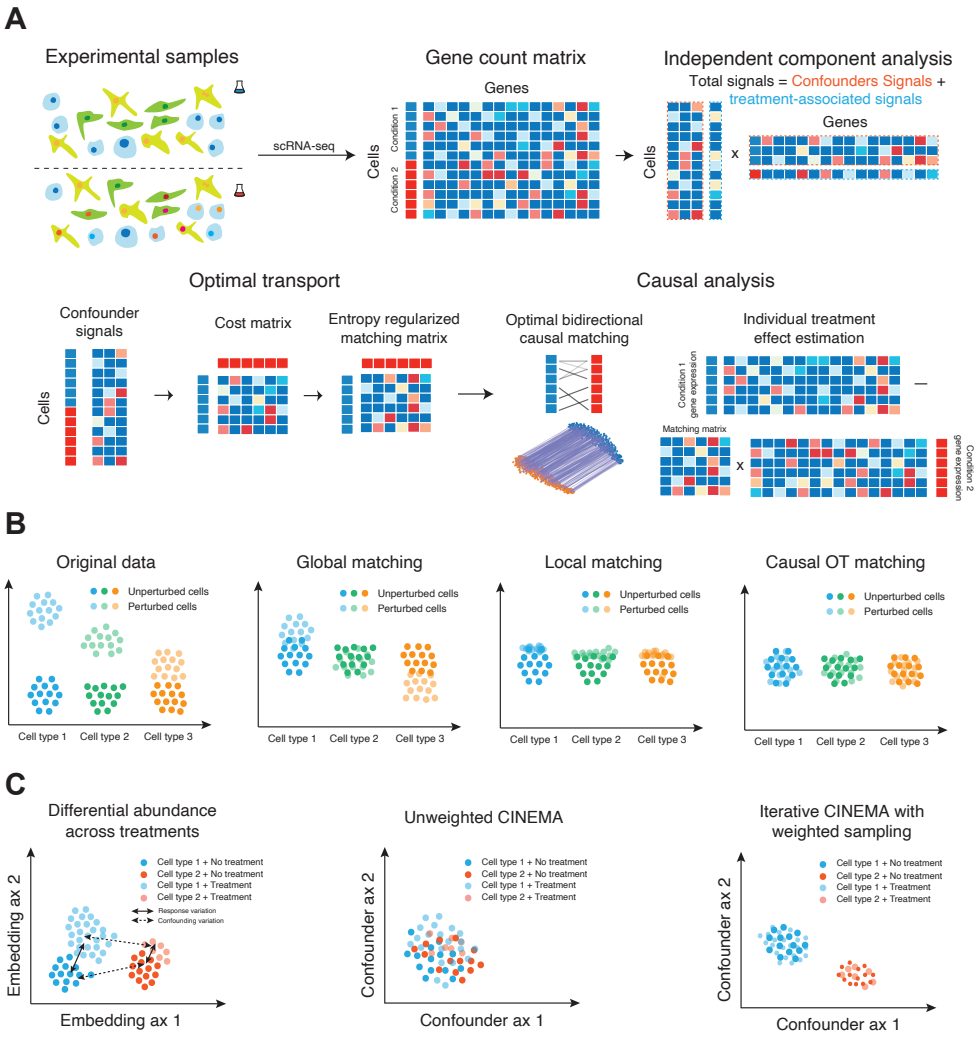

Figure 2: Overview of the CINEMA-OT framework. **A.** scRNA-seq count data is first decomposed into confounder variation and treatment-associated variation using ICA. Cells are then matched across treatment conditions by entropy-regularized optimal transport in the confounder space to generate a causal matching plan. The smooth matching map can then be used to estimate individual treatment effects. **B.** Illustration of the properties of casual OT matching compared to other common matching schemes. Global matching may have poor performance when there are confounder-specific heterogenous responses to treatment. Local matching may be susceptible to boundary effects. By contrast, CINEMA-OT balances these concerns by enforcing preservation of distributional mass. **C.** Differential abundance effects may cause spurious matching by CINEMA-OT. When such effects are present, iterative reweighting may be used to balance cell populations and learn true underlying confounding signals.

2.4 Validation of CINEMA-OT using simulated ground truth datasets

There are a number of existing methods that perform single-cell level perturbation effect analysis [3–7, 18–20, 32]. Extended figure 1 comprises a summary table of currently available methods and their capabilities. To investigate how CINEMA-OT differs from these methods we perform extensive benchmarking on a number of tasks.

We systematically compare methods including Mixscape, Harmony-Mixscape, classical optimal transport, to CINEMA-OT (with and without abundance reweighting). Our comparison is based on three categories of metrics:

1. *Direct validation of individual treatment effects.* For data sets with a ground truth, we can directly compare the estimated individual treatment effect against the true individual treatment effect for each cell. We note that these metrics can only be performed on data sets with a ground truth, such as our simulated data. These metrics include: ITE Spearman and ITE Pearson correlation.
2. *Cell distribution equalization after treatment effect removal.* In data sets without ground truth, we can measure the validity of treatment effects by examining cell population distributions after treatment effect removal. After removal, these distributions should be equivalent, subject to random noise. These metrics include average silhouette width (ASW), PC regression score (PCR), and graph connectivity.
3. *Biological effect preservation after treatment effect removal.* While removal of treatment effects should render cell distributions equivalent with and without treatment, biologically meaningful information (such as cell type and cell trajectories) should be preserved. These metrics include diffusion map-based nonparametric state coefficients and trajectory coefficients.

To obtain data with ground truth, we simulate data using the R package Splatter and Python package Scsim. Splatter is a popular package [33] for single-cell *in silico* data generation, and Scsim is a python implementation of the Splatter framework with additional support for simulating gene regulation programs as trajectories. [34] The genes in our simulated data are separated into three subsets, corresponding to the underlying trajectory, cell types, and treatment-associated genes respectively. Both the trajectories and cell type for each cell have a random signal applied to the treatment-associated genes to account for the possible confounder effects on treatment-associated genes (Figure 3B).

We have tested four scenarios: I, II: The confounders have an additive effect on treatment-associated genes, with or without confounder imbalance; III: the treatment interacts with confounders, creating confounder-specific treatment effects; IV: heterogeneity of treatment effects are caused by latent factors and are not associated with explicit confounders. For each scenario, we simulate 15 datasets.

A comparison/evaluation of different methods applied to these data shows that CINEMA-OT best estimates the individual level treatment effect in all settings considered. Moreover, CINEMA-OT performs best in removing treatment-associated effects in matching and preserves trajectories better than other methods while achieving comparable performance in cell-type preservation in the first three scenarios (Figure 3C). In the fourth scenario, we examine latent factor-specific treatment effects. Here, explicit confounder signals are uncorrelated with the treatment effects. We expect that in this scenario, the state and trajectory coefficients will be higher when treatment effects are incorrectly attributed to confounders. We see this with Mixscape, Harmony/Mixscape and classical OT (Figure 3C). Taken together, these data show that CINEMA-OT performs equivalent to or better than other methods in all tested scenarios.

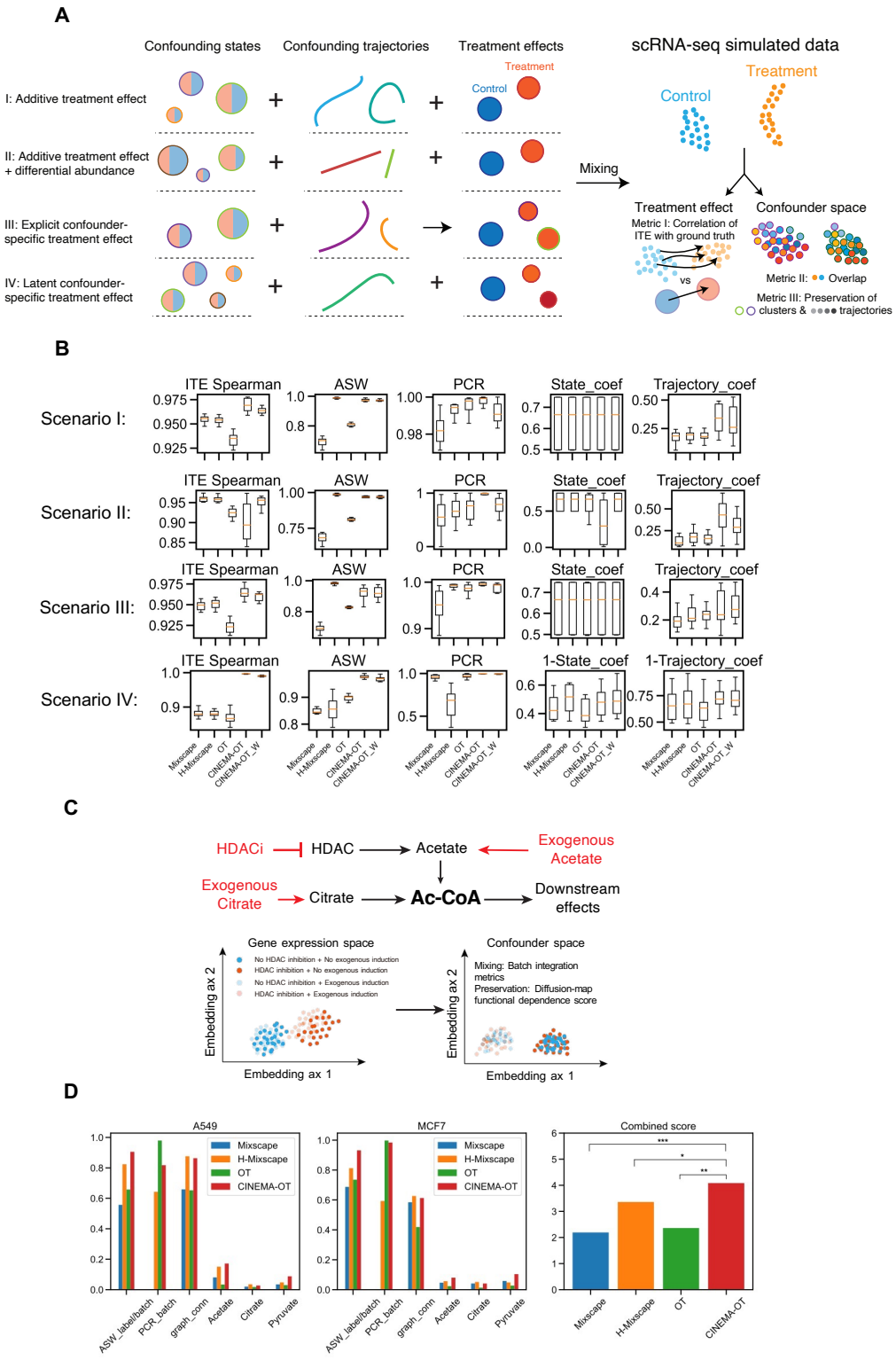

Figure 3: Benchmarking of CINEMA-OT against other methods for single-cell perturbation analysis. **A.** We simulate binary perturbation data with confounding effects using Splatter and scSim for four scenarios. We evaluate CINEMA-OT against other methods using three classes of metrics: (I) direct correlation of ITE with ground truth, (II) mixing of samples in confounder space (III) preservation of confounder signals including clusters and trajectories. **B.** Quantification of different validation metrics on synthetic data for CINEMA-OT and comparable methods. **C.** CINEMA-OT validation on HDAC-inhibitor perturbed data. Exogenous metabolites serve as confounding factors affecting HDAC inhibitor perturbation. Methods are evaluated by distributional matching across treatments and confounding factor preservation respectively. **D.** Results of the evaluation of CINEMA-OT in the A549 and MCF7 cell lines perturbed by an HDAC inhibitor. CINEMA-OT demonstrates superior performance when compared to other methods shown by aggregated rank comparisons. ITE: individual treatment effect; ASW: average silhouette width; PCR: principal components regression score; State_coef: diffusion map-based state coefficient; Trajectory_coef: diffusion map-based trajectory coefficient. Stars indicate statistical significance as follows: * ($p < 0.1$); ** ($p < 0.01$); *** ($p < 0.001$).

2.5 Validation of CINEMA-OT on real data

In order to evaluate the performance of CINEMA-OT on real scRNA-seq data, we use the sci-Plex4 single-cell drug perturbation dataset [8], which measures the response of the A549 and MCF7 cell lines to perturbation with 17 drugs. In real scRNA-seq data, we do not know the ground truth perturbation effects or ground truth confounders at the single-cell level as in simulated data. Therefore, only the unsupervised metrics for perturbation effect elimination can be still adopted precisely in this case.

Here we investigate the response to perturbation with Pracinostat, a histone deacytelase (HDAC) inhibitor, with the combinatorial induction of exogenous acetate, citrate, and pyruvate. HDAC inhibitors act as antitumoral agents through antagonizing the pro-transcriptional effects of histone acetylation and silencing the expression of oncogenic factors through chromatin remodeling [35]. As HDAC inhibitors act partly through the deprivation of Ac-CoA, we expect that the relative abundance of Ac-CoA precursors within a cell would modulate the effect of HDAC inhibitor exposure, and Ac-CoA precursors can be considered confounders [8] (Figure 3D). Particularly, this experiment coincides with the third scenario in the previous benchmark with synthetic data. Upon examination, we find that CINEMA-OT can both match the distributions across treatment conditions and preserve the confounding labels well, giving the best overall performance among the tested methods (Figure 3E, see Methods for additional details).

Cell cycle stage is a well studied confounding covariate for experimental perturbation effects. However, in Sci-plex data, we cannot evaluate the accuracy of cell cycle stage preservation since HDAC inhibition induces a G2/M cell cycle arrest [8]. As a result, to evaluate cell cycle preservation for the different methods, we quantify cell cycle information preservation after matching cells across cell lines. In order to get the cell cycle signal, we use the Tricycle package [36], which returns a vector with each entry value between 0 and 2π, representing a cell's phase in the cell cycle. We visualize cell cycle information preservation after matching cells across cell lines. CINEMA-OT is the only method that successfully matched the cells between conditions while largely preserving the cell cycle information (Extended Figure 4).

2.6 CINEMA-OT identifies heterogeneous response patterns and synergistic effects of cigarette smoke exposure in Rhinovirus infection

In order to demonstrate CINEMA-OT's ability to perform single-cell level experimental perturbation analysis, we have applied CINEMA-OT to newly-collected scRNA-seq data of rhinovirus infection in primary human airway organoids [37] (Figure 4A). The experiment comprises 4 conditions, corresponding to all combinations of cigarette smoke extract (CSE) and rhinovirus (RV) infection (mock, CSE, RV, RVCSE). The goal of this study is to probe cellular defense responses to viral infection from each airway epithelial cell type in the presence or absence of an environmental insult, for example cigarette smoke. Viral infection occurred only in a small percentage of cells but caused global expression of interferon stimulated genes (ISGs), both with or without cigarette smoke exposure. Previous studies only considered gene expression at the cluster level [38] and are unable to detect possibly heterogenous response patterns within clusters, which may be of biological and clinical relevance to the treatment of respiratory virus infections.

We first performed CINEMA-OT analysis for the RV and mock conditions (Figure 4B). We performed individual treatment effect estimation to obtain a per-cell response matrix, which we visualized and clustered (Figure 4C). By analyzing these response-based clusters, we found distinguishable sub-clusters in ciliated cells with different induction levels of ISGs (Figure 4D). The result is

further supported by visualization of typical ISGs at the sub-cluster level, including BST2, MX1, IFITM3 and others, which are known to have direct anti-viral functions [39, 40] (Figure 4E). This finding highlights the heterogeneous response patterns within a cell type and the need for sub-cluster level analysis of data in order to reveal such variations.

To further attribute different response patterns to either explicit or latent confounding variation, we performed coarse-grained analysis of the optimal transport matching matrix (Figure 4F). This approach reveals that two branches of ciliated cells in the mock condition drive the difference levels of interferon responses in the infected condition across sub-populations (Figure 4G). The identified patterns may correspond to unknown subpopulations with defense specializations or different susceptibilities to virus that may be important for further investigation.

After analysis of the effect of smoking and viral infection in isolation, synergy between these two insults was assessed by comparing the difference between infection with and without cigarette smoke exposure. Our analysis highlights the synergistic response of ISGs in the perturbation. ISGs are induced to a higher expression level in the virus infected organoid without exposure to cigarette smoke, compared to the level in the infected organoid exposed to cigarette smoke (Figure 4H). The result is consistent with previous works that use traditional approaches to identify relationships between cigarette smoke exposure and immune response to virus infection, which have shown that airway epithelial cells have diminished innate immunity response in smokers [41–44].

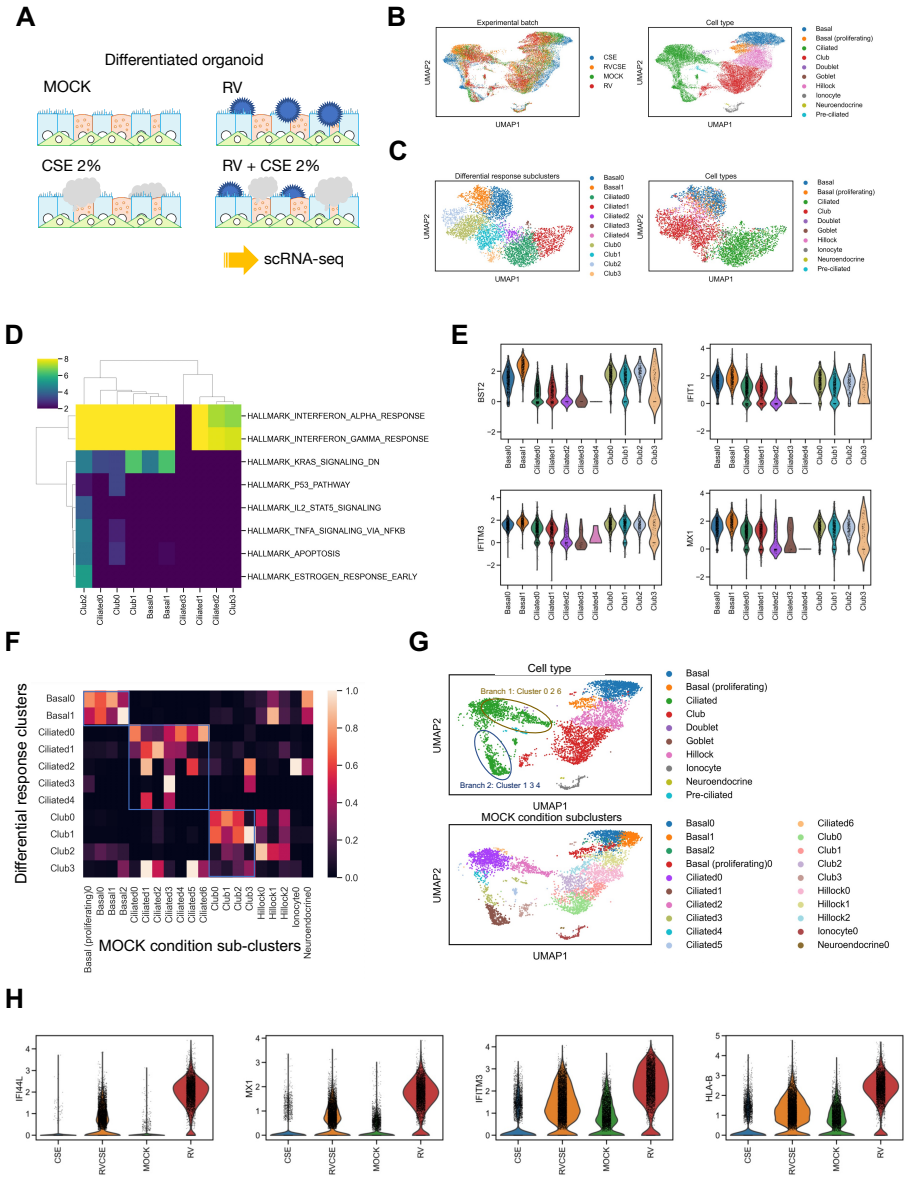

Figure 4: CINEMA-OT identifies heterogeneous defense response of human airway epithelial cells to rhinovirus and cigarette smoke extract. **A.** Overview of experimental design. Differentiated airway epithelial organoids are challenged with mock (control) or rhinovirus 1A infection (RV), with or without cigarette smoke extract (CSE) exposure. **B.** UMAP projection of expression data labeled by perturbations and cell types. **C.** UMAP projection of the individual treatment effect matrix obtained by CINEMA-OT from the RV response without CSE exposure, colored by response cluster and cell type. **D.** Gene set enrichment analysis of response clusters identified by CINEMA-OT. **E.** Violin plots of several representative interferon stimulated genes (ISGs) in differential response clusters. **F.** Coarse-grained matching matrix visualization. The horizontal axis represents clusters in the control condition. The vertical axis represents differential response clusters. **G.** UMAP projection of control condition expression data, colored by cell types and cell subclusters. Two branches of ciliated cells with correspondence to different levels of immune responses are highlighted. **H.** Violin plots of representative ISGs where CINEMA-OT identifies synergy between CSE exposure and rhinovirus infection.

2.7 CINEMA-OT reveals principles of innate immune response modulation from combinatorial interferon stimulation

Type I, type II, and type III interferons (IFNs) act as central regulators of immune responses during intracellular pathogen infection, cancer and in auto-immunity. However, despite the identification and adoption within the literature of a core set of interferon-stimulated genes (ISGs), IFN responses can vary widely by cell type, by individual, by IFN stimulus type, by chronicity of exposure, and by combination with signals delivered by other cytokines. In other words, the interferon response is highly context dependent. This complexity, heterogeneity and context-specificity of IFN signaling can lead to counter-intuitive results. For example, IFNg is proposed to play both stimulatory and suppressive roles in cancer, and type I IFNs are used both as an immunosuppressant to treat multiple sclerosis and as immunostimulatory adjuvant treatments for cancer (e.g. melanoma) and chronic viral infection (e.g. HCV). To model the complexity of IFN signaling, we conducted acute (2 days) and chronic (7 days) stimulations of peripheral blood immune cells from multiple healthy donors with type I, type II and type III IFNs, separately as well as in combination with other cytokines such as TNFa and IL-6. Using CINEMA-OT, we sought to map the determinants of IFN response by IFN type, timing, cell type, and combination with other cytokines that can be used to decode IFN effects on immune cells in diverse biological contexts. Specifically, we sought to distinguish the effects of type I, type II, and type III IFN from one another and to identify synergistic effects of IFN combinations with other cytokines. In the following analyses we focus on a single donor as an illustration.

In order to understand the underlying structure of PBMC cellular response to interferon stimulation, we use CINEMA-OT to match treatment conditions to the untreated (control) condition. This analysis highlights the underlying hierarchical structure of cellular responses. Visual inspection of the UMAP projections of the response space show that in acute stimulation type I, type II, and type III IFN responses cluster separately. IFNb in combination with IFNg also occupies its own cluster after 2 days of stimulation. At day 7, responses to type I, type II, and type III interferon alone remain distinct from each other, but notably, type I IFN in combination with type II IFN (IFNb + IFNg) appears to co-cluster with type I IFN responses, suggesting a modulatory effect of chronic stimulation on this combination of signals (Figure 5C).

Next, we focus on analyzing the treatment effects of IFNb. CINEMA-OT analysis highlights the induction of coordinated immune responses across cell types as well as cell-type specific responses. For example, at day 2, despite a global change in ISGs in most cell types, a sub-cluster of monocytes demonstrates a unique program characterized by decreased APOBEC3A and MARCKS expression compared to other monocytes. Further CINEMA-OT attribution analysis suggests this sub-cluster corresponds to a specific sub-cluster of monocytes marked by increased IDO1 expression prior to treatment (Figure 5 D-F).

To estimate the synergistic effects of combinatorial cytokine stimulations at day 2, we used CINEMA-OT to calculate the synergy score per gene by cell pair. We next performed gene synergy score analysis by computing the gene-wise synergy score (See Methods). The gene synergy score analysis identified genes that were synergistically induced by each combinatorial perturbation (Figure 5G), including ACP5, APOC1, APOE, CCL3L1, CD9, GPNMB, TREM2, APOBEC3A, CCL8, IL1RN, CHI3L1, CXCL9, and CCL7.

Based on selected significant synergy genes, we are able to summarize the cell wise synergy effect by taking the norm over selected synergy genes. We have found that monocytes exhibit the most significant synergistic regulation compared with other cell types (Figure 5H). In monocytes, a subset of significant synergistic genes were expressed in the control condition and were not expressed in conditions with IFNb or IFNg present, including APOC1, APOE, CCL3L1, CD9, GPNMB, TREM2, and ACP5 (Figure 5I). APOBEC3A, CCL8 and IL1RN were expressed at a higher level in combinatorial treatment conditions, while CHI3L1, CXCL9, and CCL7 have specific expression that was induced by only a minority of combinations (Figure 5I). Finally, we visualize the UMAP projection of the synergy matrix while highlighting cell-wise overall synergy effect and several specific synergistic genes (Figure 5I). Our analysis of the data with CINEMA-OT sheds light on the underlying biological regulation and mechanisms of the heterogeneous cellular response to interferon stimulation. Particularly, the synergy analysis by CINEMA-OT provides important insights for future investigation.

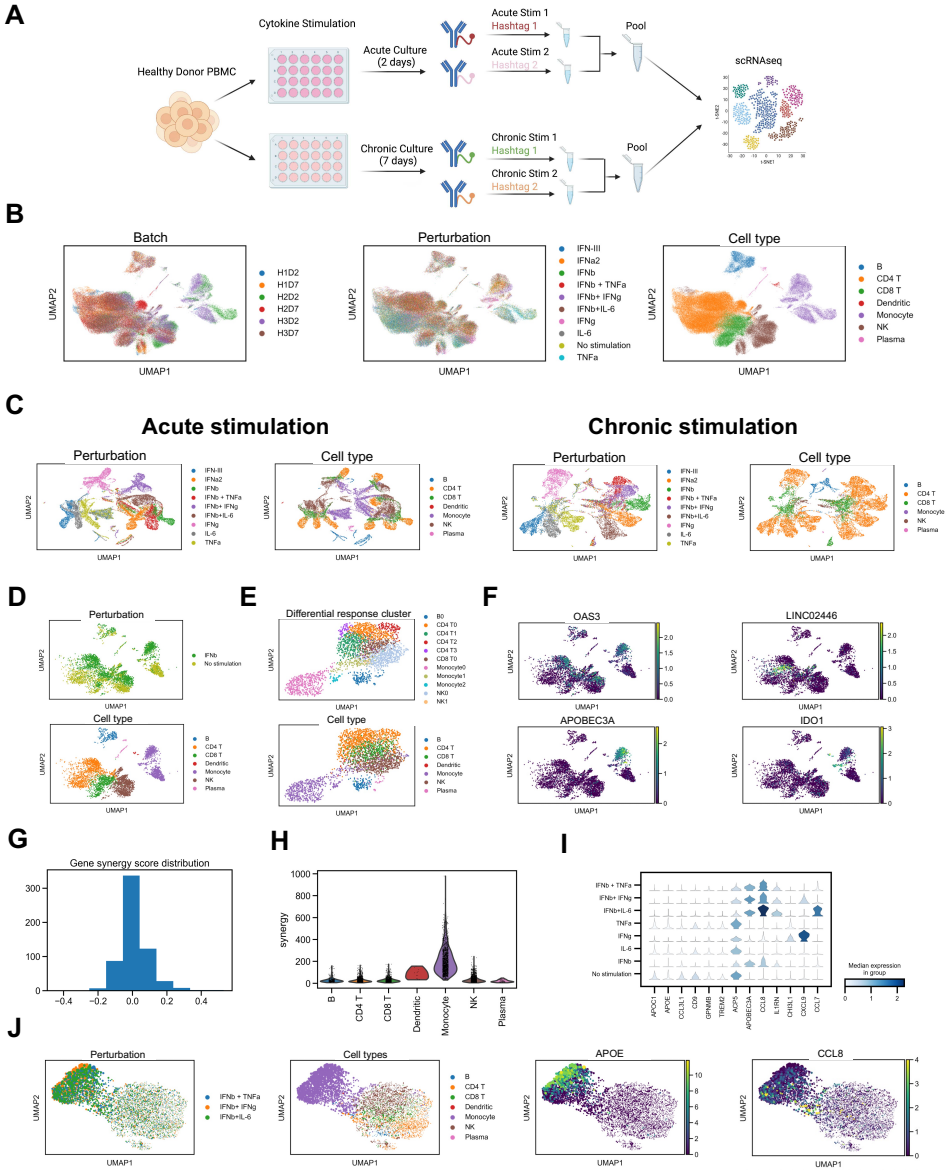

Figure 5: CINEMA-OT reveals combinatorial mechanics of acute and chronic cytokine stimulation. **A.** Illustration of experimental design. **B.** UMAP projection of expression data colored by samples, perturbations, and cell types. In sample labels, H refers to the donor number, and D refers to the number of days of stimulation. **C.** UMAP projection of the individual treatment effect matrices from CINEMA-OT across acute stimulation (day 2) and chronic stimulation (day 7) for donor 3. Projections are colored by cytokine stimulation and cell type. **D.** UMAP projection of expression data after acute (2 day) stimulation with interferon beta in donor 3, colored by perturbation and cell types. **E.** UMAP projection of individual treatment effects identified by CINEMA-OT after acute stimulation with interferon beta, colored by response cluster and cell type. **F.** UMAP projection of the expression data highlighting selected genes showing markers of cellular activation and inhibition. **G.** Gene synergy score distribution combining IFNb + TNFa, IFNb + IFNg, and IFNb+IL-6 combinatorial treatment for donor 3 after 2 days of stimulation. **H.** Violin plot of the cell synergy scores per cell type computed over only statistically significantly synergistic genes. **I.** Stacked violin plot showing the expression of significantly synergistic genes per stimulation condition in monocytes for donor 3. **J.** UMAP projection of the synergy matrix from CINEMA-OT across three conditions. Dot size indicates cell synergy scores which have been normalized to the minimum and maximum values for all conditions visualized. APOE and CCL8 are selected as representatives to show two modes of synergistic responses in monocytes.

3 Discussion

With fast developing high-throughput screening technologies and rising numbers of datasets, single-cell level experimental effect analysis is becoming a critically important task. Current analytical approaches aiming to tackle this task suffer from multiple fundamental challenges. For example, differential abundance methods estimate the local likelihood of treatment variables in order to reveal perturbation effects but cannot integrate confounder signals and cannot offer direction on how the perturbation can shift the gene expression distribution. Meanwhile, approaches using local matching such as k-nearest neighbors are vulnerable to outliers or subclusters caused by latent confounders, such as perturbation escape. More complex methods that use a large parameter space may suffer from overfitting and uninterpretability.

With CINEMA-OT, we are able to overcome the aforementioned challenges with a causal framework and our corresponding scalable and accurate algorithm. In this study, we applied CINEMA-OT to deconvolve confounder effects due to exogenous addition of metabolites (Figure 3), cigarette smoke exposure and viral infection (Figure 4) and combinatorial cytokine treatments of differing durations (Figure 5). In each case we were able to separate confounder and treatment effects as well as identify synergies among combinatorial treatment effects.

Although examples in the text mainly emphasize experiments with binary treatment conditions (e.g. treated or untreated), CINEMA-OT may also function with continuous perturbations (e.g. continuous treatment time or graded treatment dose). In fact, by using a functional dependence metric instead of a binary statistical test, causal matching in CINEMA-OT may be naturally extended to a multi-value perturbation design.

Two potential challenges for CINEMA-OT can arise due to bias-variance tradeoffs in optimal transport and the choice of dimensionality for CINEMA-OT's internal representation. For the first challenge, a large smoothness threshold in the entropy regularized method can overly smooth the obtained matching map and cause false positives by incorrectly identifying confounder variation as treatment-associated variation. However, too small a threshold would both harm the method's stability and cause high variance. In practice, an adequate threshold can be chosen based on coarse-graining of the matching matrix. For the second challenge, we have applied a rank estimation procedure to quantify the instrinsic dimensionality of the count matrix [45]. For larger datasets, this procedure can be computationally expensive and we have found that a fixed rank of 20 generally yields good performance.

CINEMA-OT is a tool designed for the estimation of gene signatures causally associated with treatment responses. While we have implemented an iterative reweighting procedure to account for differential confounder abundance that may arise in response to treatment, CINEMA-OT is not designed for cases where large-scale changes to confounder distributions are the primary effects of interest, such as proliferation or cell death. In those cases, tools such as MELD, MILO, or DA-seq may be more suitable [17, 14, 15].

Although several works have explored single-cell level perturbation analysis, none of the methods to date are based on a strict causal framework. CINEMA-OT provides the first causally-aware approach to systematically characterize treatment-associated effects in single-cell data. We anticipate that CINEMA-OT will be widely adopted in single-cell perturbation analysis.

Methods

CINEMA-OT

CINEMA-OT is an unsupervised method for separating confounding signals from perturbation signals for matching cells via imputing counterfactuals and computing perturbation effect at a single-cell level (`https://github.com/vandijklab/CINEMA-OT`). The detailed workflow of CINEMA-OT is as follows.

1. **Causal framework and formal proof**

   Here we give an rigorous treatment of the causal framework and underlying assumptions in CINEMA-OT.

   Assumption 1: Independent sources and noise. *Confounding factors and treatment events are pairwise independent random variables.*

   Assumption 2: Linearity of source signal combinations. *Confounding gene signatures can be modeled as a linear combination of confounding sources plus an independent noise term. The measured gene signatures can be modeled as arbitrary functions of confounding factors and treatment events plus an independent noise term.*

If we further assume the effects of confounders and treatments on perturbation-associated gene signatures are additive, then we may consider the problem of separating confounder gene signatures from treatment-associated gene signatures as equivalent to solving the blind source separation problem. However, we can no longer reveal the potential interactions between confounders and treatments in this case. Here we take a different strategy by proving the identifiability of confounding factors in our framework.

Assume the gene count matrix is a gene by cell matrix X, the corresponding principal component matrix is $\hat{X} \in \mathbb{R}^{m \times n}$, the sources are $S = \{s_i\}_{i=1,...,l}$, each source is a vector with size the number of cells and $F = \{f_i\}_{i=1,...,m}$ is a set of arbitrary functions. Then the data generation mechanism is given by

$$\hat{X} = F(S, z) = \begin{bmatrix} f_1(s_1, ..., s_l, z) \\ \vdots \\ f_m(s_1, ..., s_l, z) \end{bmatrix}$$

where z is the treatment indicator.

With Assumption 2, $\hat{X}$ can be defined as a combination of confounding factors and perturbation-associated factors up to arbitrary invertible linear transformations A and B, and where $g_1, ..., g_{m-l}$ are arbitrary functions:

$$\hat{X} = B \begin{bmatrix} AS \\ g_1(s_1, ..., s_l, z) \\ \vdots \\ g_{m-l}(s_1, ..., s_l, z) \end{bmatrix}$$

Then upon any matrix factorization, applying a functional dependence metric, we are able to distinguish if signals are dependent on the treatment variable z. Therefore we have

$$\hat{X} = C\hat{S},$$

where $\hat{f}_1, ...\hat{f}_k$ and $\hat{g}_1, ..., \hat{g}_k$ are defined by:

$$\hat{S} = \begin{bmatrix} \hat{f}_1(s_1, ..., s_l) \\ \vdots \\ \hat{f}_k(s_1, ..., s_l) \\ \hat{g}_1(s_1, ..., s_l, z) \\ \vdots \\ \hat{g}_{m-k}(s_1, ..., s_l, z) \end{bmatrix} = C^{-1} B \begin{bmatrix} AS \\ g_1(s_1, ..., s_l, z) \\ \vdots \\ g_{m-l}(s_1, ..., s_l, z) \end{bmatrix}.$$

We define $\hat{A} = B^{-1}C$, in which case we have

$$S = A^{-1}(\hat{A}\hat{S})_{1:l} = A^{-1}(\sum_j \hat{A}_{ij}\hat{S}_j)_{i=1:l}.$$

Because $\hat{g}$s are pairwise independent functions of z, $\hat{A}_{ij}$ has to be zero for $j > k$, otherwise $A^{-1}(\sum_j \hat{A}_{ij}\hat{S}_j)_{i=1:n}$ would be a function of z, leading to a contradiction.

Therefore we have

$$S = A^{-1}(\sum_{j=1}^{k} \hat{A}_{ij}\hat{S}_j)_{i=1:l} = A^{-1}(\sum_{j=1}^{k} \hat{A}_{ij}\hat{f}_j)_{i=1:l} = A^{-1}\hat{A}^*\hat{f}.$$

Here $\hat{A}^*$ is a $l \times k$ matrix. Therefore because S is of rank l, we must have $k \geq l$, and $\hat{f}$ is a linear transformation of S.

If $k > l$, then because $\hat{f}$ defines a linear transformation on S, $k > l$ indicates $\hat{f}$ is not of full rank. Therefore matrix $\hat{S}$ is not of full rank, contradicting with its a invertible transformation of a full-rank signal matrix. The contradiction leads to the only possibility of $k = l$.

With $k = n$, because A and $\hat{A}^*$ (the sub-matrix of the full-rank matrix $\hat{A}$) are both invertible, according to ICA identifiability theorem, we have

$$S = \hat{S}_{1:l} = (C^{-1}\hat{X})_{1:l}$$

up to a permutation.

In summary, we have proved that upon correctly identifying $\hat{f}$s in our framework, we can identify all underlying confounders up to a permutation, even if there are additional nonlinear treatment-associated signals. The key difficulty we have overcomed is $g_i(s_1, s_2, ..., s_l, z)$ are generally dependent across different is, therefore the ICA identifiability theorem cannot be applied. Note that we cannot guarantee the recovery of treatment-associated signals here, which is why after identifying the confounders, we applied a non-parametric causal matching approach to fully reveal underlying causal processes in the data.

We cannot guarantee theoretically that our threshold procedure selects correct confounders and treatment-associated signals. Accounting for the possible uncertainty, it can still be seen that the matching matrix constructed by CINEMA-OT smoothly interpolates between the matching matrix of Mixscape and null matching matrix used in single-cell differential expression analysis.

2. **Rank initialization**

   In order to perform CINEMA-OT, we first need to initialize the expected matrix rank, representing the total signal number. We here offer two possible approaches for rank initialization in CINEMA-OT.

   Biwhitening [45] is a recently-developed method to remove independent heteroskedastic noise in data with inspirations from random matrix theory. It does diagonal matrix transformation of the data on both sides and thresholding based on the Marchenko-Pastur law [46]. After thresholding, we can get the true matrix rank and the matrix's low dimensional approximation. Mathematical details of biwhitening can be seen in [45]. In CINEMA-OT, we have implemented a version of biwhitening with fixed hyperparameters.

   In large datasets, we suggest using prespecified rank values. Empirically, we have found that CINEMA-OT is robust to rank selection at certain ranges and can give a good performance when $\mathrm{DimSize} = 20$.

3. **Signal selection with independent component analysis**

   Independent component analysis is already a well-addressed method in data analysis and has various implementations. Here we use the FastICA implementation from the package `sklearn.decomposition` [47]. Prior to FastICA, input data is PCA-transformed using scanpy [48].

   In order to identify confounder signals and treatment-associated signals, we have adopted a recently proposed cross rank coefficient [28], which is able to quantify the functional dependence between ICA signals and query signals (in this case, the treatment signals). We use the implementation of this method from the XICOR package in R. The threshold of the cross rank coefficient is set to be 0.5 to 0.75 in this study. We note that tuning the threshold parameter has a practical meaning in the algorithm. High thresholds correspond to less tolerance for false positive treatment signals, which leads to local matching more similar to Mixscape analyses. Meanwhile, setting a low threshold means less tolerance for false positive confounder signals and can lead to lower resolution of matching, which, in the extreme case, coincides with single-cell differential expression testing methods.

4. **Optimal transport matching**

   After selecting confounding signals, we perform matching across treatments via optimal transport, which provides a smooth transport map and does not require neighbor number selection. Here we consider the entropy regularized optimal transport formulation, which can be efficiently solved by the Sinkhorn-Knopp algorithm [30]. In this formulation of the problem, the penalty coefficient act as a hyper parameter influencing the resolution and smoothness of the transport map. We have empirically determined that the optimal value for the penalty coefficient often lies within the range $(10^{-6}, 10^{-3})$ multiplied by the number of confounding signals.

Algorithm 1 CINEMA-OT

Require: Count matrix $X_0 \in R^{m \times n}$, treatment vector $z \in \{0,1\}^n$, dimension size r, signal filtering threshold d, smoothness s.

$\text{DimSize} \leftarrow r, \text{Thres} \leftarrow d, X \leftarrow X_0$.
unmixing matrix B, source matrix $S \leftarrow \text{ICA}(X, \text{DimSize})$;
$c \leftarrow \text{zeros}(\text{DimSize})$
for $i = 1 : \text{DimSize}$ **do**
    $c_i \leftarrow \text{xicor}(S[i,:], z)$; ▷ Compute Chatterjee cross rank coefficient
end for
$S^c \leftarrow S[c < \text{Thres}, :]$ ▷ Thresholding to separate confounder signals S^c
$M \leftarrow \text{OT}(S^c[:, z = 0], S^c[:, z = 1], \text{smoothness} = s * S^c.\text{shape}[0])$ ▷ M: Matching matrix
$D \leftarrow X_0[:, z = 1]M - X_0[:, z = 0]$ ▷ ITE matrix computation
Downstream analysis.

Algorithm 2 OT

Require: Confounder signals S_1, S_2, weights $w_1 = \text{None}, w_2 = \text{None}$, smoothness s.

if w_1 is None **then**
    $r \leftarrow 1/S_1.\text{shape}[0], c \leftarrow 1/S_2.\text{shape}[0]$
else
    $r \leftarrow w_1/w_1.\text{shape}, c \leftarrow w_2/w_2.\text{shape}$
end if
$D \leftarrow \text{PairwiseDistance}(S_1, S_2)$.
$A \leftarrow \exp(-D * D/s)$ ▷ Elementwise multiplication for D here
$M = \text{SinkhornKnopp}(A, \text{setr} = r, \text{setc} = c)$ ▷ Sinkhorn-Knopp algorithm
return M

Iterative reweighting CINEMA-OT

If confounder signals are not independent of treatment indicators, as in the case of differential abundance, confounder signals and treatment-associated signals may not be completely unmixed by ICA. Even if we select the confounder signals by thresholding as previously described, there may still remain treatment signal within the confounder signals. We have implemented a heuristic approach to enable better separation of these signals through iterative application of ICA and reweighting. If the confounders are independent of treatment labels, then the local abundance of cells in the confounder space should be balanced. If the cells are not balanced in the confounder space, then we attempt to impose balance on the space by reweighting the cells.

In our implementation, we use MELD [17] to estimate each cell's treatment label likelihood. The likelihood is then used for estimation of cell-wise weights. In our case, we use the overlap weights for numerical stability, as the method can assign low weights to cells with similar treatment label neighborhoods.

We note, however, that imbalance in the confounder space may also occur when the threshold parameter d is set too high, causing misidentification of treatment-associated factors as confounding factors. CINEMA-OT has no way of distinguishing between these two scenarios, but our benchmarking suggests that the iterative reweighting procedure assigns correct weights in practice, achieving good separation of confounding and treatment-associated variation. In addition, CINEMA-OT offers users the ability to specify known confounder labels (e.g. cell type, cell cycle), which may be directly used for balancing as an alternative to MELD, without the need for an iterative procedure.

Algorithm 3 Iterative reweighting CINEMA-OT

Require: Count matrix $X_0 \in R^{m \times n}$, treatment vector $z \in \{0,1\}^n$, dimension size r, signal filtering threshold d, smoothness s.

DimSize $\leftarrow r$, Thres $\leftarrow d$, $X \leftarrow X_0$.
while not converge **do**
  unmixing matrix B, source matrix $S \leftarrow \text{ICA}(X, \text{DimSize})$;
  $c \leftarrow \text{zeros}(\text{DimSize})$
  for $i = 1 : \text{DimSize}$ **do**
    $c_i \leftarrow \text{xicor}(S[i,:], z)$; ▷ Compute Chatterjee cross rank coefficient
  end for
  $S^c \leftarrow S[c < \text{Thres}, :]$ ▷ Thresholding to separate confounder signals S^c
  $p \leftarrow \text{MELD}(\text{data} = S^c, \text{SampleLabel} = z)$.
  $w[z=0] \leftarrow p[z=0,1], w[z=1] \leftarrow p[z=1,0]$. ▷ Compute propensity score
  $X \leftarrow \text{sample}(\text{size} = 2n, \text{data} = X, \text{Weight} = w)$
end while
$\hat{S} \leftarrow (BX_0)^T$
$\hat{S}^c \leftarrow \hat{S}[c < \text{Thres}, :]$
$M \leftarrow \text{OT}(\hat{S}^c[:, z=0], \hat{S}^c[:, z=1], w_0 = w[z=0], w_1 = w[z=1], \text{smoothness} = s)$ ▷ M: Matching matrix
$D \leftarrow X_0[:, z=1]M - X_0[:, z=0]$ ▷ ITE matrix computation
Downstream analysis.

Downstream analysis

1. **Visualization and clustering of the ITE matrix**

   With the ITE matrix computed by matching counterfactuals, we are able to numerous standard analyses. We may employ dimensionality reduction techniques such as t-SNE, UMAP, or PHATE [49–51] to visualize clusters in the response space. We may also employ clustering techniques, such as Leiden clustering [23] to group cells by similarity of treatment responses.

2. **Synergy analysis**

   For synergy effect, we compare ITE matrices for two treatment conditions against the ITE matrix for the combined treatment. Formal derivation of the synergy score is given as follows.

   Consider $D_{A=1,B=0}$ as the ITE matrix for treatment A alone, $D_{A=0,B=1}$ as the ITE matrix for treatment B alone, and $D_{A=1,B=1}$ as the ITE matrix for the combined treatment. We may define a synergy matrix Ψ as:

   $$\Psi = D_{A=1,B=1} - (D_{A=1,B=0} + D_{A=0,B=1})$$

   Where each entry $\Psi_{g,c}$ represents the synergy score for gene g and cell c. In order to test if a particular gene g has significant synergistic effect, we formulate the problem as if we should reject

   $$H_0 : E(\Psi_{g,c}) = 0, \ \forall c.$$

   Note here if we apply no normalization, we are aiming for additive synergy; if we instead apply log normalized data, H_0 would test for multiplicative synergy.

   We assume that different cells are unlikely to have opposite synergy effects, allowing us to relax H_0 as:

   $$H_0 : E(\overline{\Psi}_{g,:}) = 0.$$

   Assume the new H_0 holds, then for each gene g, we compute the empirical synergy:

   $$\overline{\Psi}_{g,:} = E(\overline{\Psi}_{g,:}) + \epsilon_1 + \epsilon_2 - \epsilon_3 - \epsilon_4.$$

   Because here H_0 holds, the expectation of the noise term is zero.

Assume the noise is Poisson, then with the property of Poisson distribution, in the case of log normalization, ϵ_is are averages of i.i.d. scaled log1p Poisson distribution with zero expectation. With the delta method, the variance of the noise term is approximated as:

$$\mathrm{Var}(\epsilon_1+\epsilon_2-\epsilon_3-\epsilon_4)=\frac{1}{n^2}\sum_{c=1}^{n}(\frac{\lambda_{1c}}{(1+\lambda_{1c})^2}+\frac{\lambda_{2c}}{(1+\lambda_{2c})^2}+\frac{\lambda_{3c}}{(1+\lambda_{3c})^2}+\frac{\lambda_{4c}}{(1+\lambda_{4c})^2}).$$

Where $\lambda_{ic}, i\in\{1,2,3,4\}$s are counterfactual cell gene expression expectation for each cell in 4 conditions.

We note the formula $\frac{\lambda}{(1+\lambda)^2}$ is self-standardized as it is a smooth function and is near zero for either large or small λs. Therefore, to simplify the statistical test, we assume

$$\mathrm{Var}(\epsilon_1+\epsilon_2-\epsilon_3-\epsilon_4)=\mathrm{const}.$$

In this case, if we define

$$\text{Synergy score}=|\overline{\Psi}_{g,:}|,$$

identifying most synergistic genes among all genes can be turned into comparing the synergy score over all genes.

3. **GSEA analysis**
   For differential gene expression significance, we have applied the non-parametric Wilcoxon signed-rank test. We apply a p-value threshold (10^{-5}) and expression fold change threshold selected by the user to identify significantly regulated genes. These genes are input into GSEApy for analysis by functional signatures [52, 53].
4. **Attribution analysis**
   By clustering cells both by treatment responses (i.e. using the ITE matrix) and control condition clusters (i.e. cell subtypes), the matching matrix from CINEMA-OT can be coarse grained. The resulting course-grained matching matrix is of shape $\mathrm{ResponseClusterNumber}\times\mathrm{ControlClusterNumber}$. Each column of the matrix gives the likelihood of a control condition cluster to have different modes of response. By reading each row of the matrix, we are able to attribute each response to the underlying control condition cluster.

Data simulation and analysis

For Splatter data simulation, we simulate 500 gene by 1000 cell count matrices with 2-4 underlying cell states. Then we simulate two orthogonal trajectories, each with 250 genes. After creating confounder genes, we next simulate 500 outcome genes from two groups, enabled by the `group` function in Splatter. Each cell has an equal probability of being assigned into outcome gene cluster 1 (group 1) and cluster 2 (group 2). In the first two scenarios, The confounders are simulated to have random linear effects on the outcome genes, represent potential mixing of confounders and outcomes in the same gene, with or without differential abundance. In the third scenario, the confounders are simulated to have both linear mixing effects and state-specific effects on treatment-associated genes. We simulate 15 times with different random seeds to generate 15 gene count matrices with size 1000×1500 (cells by genes). Then the data is normalized, log transformed, and analyzed by our implemented methods respectively.

For Scsim data simulation, we simulate 500 gene by 1000 cell count matrices with 2-5 underlying cell states with 2 gene regulation programs. We use Scsim to simulate the fourth scenario mentioned in the main text, where the heterogeneity of treatment effects is generated by random sampling independent of the confounders. We simulate 15 times with different random seeds to generate 15 gene count matrices with size 1000×1500 (cells by genes). Then the data is normalized, log transformed, and analyzed by our implemented methods respectively.

For the Mixscape analysis, we have implemented a simple version in Python that matches cells across conditions according to the descriptions in [18]. For Harmony-Mixscape analysis, we have used the Python package `harmonypy` (https: //github.com/slowkow/harmonypy) with default settings [54]. For OT analysis, we implemented a function that calls entropy-regularized optimal transport with a tunable smoothness parameter. For CINEMA-OT, we run the analysis with default settings and smoothness is set to be 5×10^{-4}.

Benchmarking metrics

In benchmarking studies, we have implemented three categories of metrics. The first method category evaluates ITE estimation accuracy.

ITE Pearson, ITE Spearman

We evaluate the Pearson / Spearman correlation between output ITE matrices and the ground truth treatment effects (derived using the ground truth matching).
The second and third method categories include intermediate metrics for evaluating batch effect removal and supervised biological effect preservation respectively.

ASW, PCR, Graph connectivity

These metrics are batch mixing metrics used to evaluate batch correction methods performance in the systematic benchmarking paper [55]. CINEMA-OT uses these metrics to evaluate mixing in confounder space, as a surrogate for correct matching that can still be measured when ground truth labels are not present. We use the implementations of these metrics from package scib [55].

Diffusion-map dependence coefficient

In order to evaluate preservation of underlying confounders, we use diffusion-map dependence coefficients. We calculate these coefficients for both cell state and cell trajectory. As our simulated data do not form a well-defined trajectory, and multiple orthogonal trajectories may be simultaneously present in the data, we approach trajectory preservation differently from [55].
We evaluate the maximum possible functional dependence coefficient, a rank-based measure, between true order and diffusion map eigenvectors. Low values of the index would indicate there are no components that are of high functional dependence with respect to the trajectory, therefore the information of the trajectory is not well preserved.
The Laplacian eigenvectors rely on kernel function selection. In the case of a covariance kernel, the method reduces to computation of the functional dependence between ground truth and confounding covariates identified by each method. In our case, we select the default setting provided by the `scanpy.tl.diffmap` implementation.
For discrete labels, we use one-hot encoding to create n vectors where n equals the number of labels. Then for each vector, we seek the maximum functional dependence score as described above, and then take average over all n coefficients to get an average score as the final output.

Sci-Plex4 data

The Sci-Plex4 data was accessed from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM4150379 with GEO accession number GSM4150379. The data is preprocessed via protocol https://github.com/manuyavuz/single-cell-analysis/blob/main/single_cell_analysis/datasets/sciplex.py. After preprocessing, we normalized and log transformed the raw count matrix and perform subsequent analysis described in main text.
After estimating all metrics, each metric is rescaled so that the sum across all methods tested equals 1. Then we sum the rescaled score over metrics and cell types to calculate the combined score. The statistical test is performed by Wilcoxon signed-rank test `scipy.stats.wilcoxon` over the rescaled metrics.

Rhinovirus infection data

Primary human bronchial epithelial cells from healthy adult donors were obtained from commercial vendor (Lonza) and cultured at air-liquid interface according to the manufacturers instructions (Stem Cell Technologies) using reduced hydrocortisone. Cells were kept at air-liquid interface for 4 weeks before experiment; maturation of beating cilia and mucus production was confirmed using light microscope. Cells were then infected with mock or 10^5 PFU human rhinovirus 1A per organoid, with or without exposure to 2% cigarette smoke extract (CSE). Single cell suspension is collected by trypsin digestion at 5 days post infection and submitted to single cell RNA sequencing using The 10X Genomics single-cell $3'$ protocol. The final dataset contains 24767 cells and 23529 genes in 4 samples (mock, RV, CSE, RVCSE). We used a standard cell marker database to annotate clusters identified by Louvain clustering of a BB-kNN graph based on Euclidean distances between cells in 50-dim PCA space. Lasso regression was used to determine which cell markers were most predictive in a one-vs.rest scheme for automated annotations. These markers are checked based on known cell type markers of airway epithelial cells [56], based on which each cell is assigned to be of one of eight cell types: proliferating basal, basal, hillock, club, goblet, pre-ciliated, ionocyte and neuroendocrine cells.
CINEMA-OT analysis on MOCK and RV was run with default parameters with `smoothness=5e-6`. Synergy analysis was performed with default parameters.

Interferon treatment data

PBMC processing and in vitro culture

The study was approved by Institutional Review Boards at Yale University (following Yale melanoma skin SPORE IRB protocol). Healthy donors consented to donation of peripheral blood for research use.

Human PBMC were isolated using Lymphoprep density gradient medium (STEMCELL). PBMC were plated at 1 million cells per ml and stimulated with 1000U/ml human IFNa2 (R&D systems), 1000U/ml human IFNb (pbl assay science 11415), 1000U/ml human IFNg (pbl assay science), 1ug/ml human IFN-III /IL-29 (R&D systems), 100ng/ml human IL-6 (NCI Biological Resources Branch Preclinical Biologics Repository), 20ng/ml human TNFa (R&D systems), and combinatorial cytokines IFNb+ IL-6, IFNb+ TNFa, IFNb+ IFNg at indicated concentrations above for up to 48 hours.

Cell enrichment and 10x sample preparation

Cultured cells were collected stained with TotalSeq anti-human hashtags C0251-C0260 (Biolegend), viability dye (zombie red, Biolegend) and anti-human CD45-FITC (clone HI30, Biolegend) and enriched for live CD45+ cells using BD FACS Aria II. Sorted cells were then resuspended to 1200 cells per ul and barcoded for multiplexed single cell sequencing using 10x Genomics 5v2 chemistry (10x Genomics, PN-1000263).

Sequencing and 10x sample alignment

Single cell RNA sequencing libraries were sequenced on Illumina NovaSeq at read length of 150bp pair-end and depth of 300 million reads per sample.

scRNA-seq data analysis

Data from three donors across Day 2 and Day 7 are concatenated together into labeled anndata objects for analysis. For each of the 6 samples, we filtered cells with less than 200 genes and we filtered genes expressed in fewer than 3 cells. For further quality control, cells with a high proportion of mitochondiral reads (> 7%) were excluded. The distribution of genes per cell was visually inspected and upper thresholds selected on a per-sample basis to exclude doublets. For each of the samples, the upper threshold was selected as [6000,3500,4000,3500,4500,3500] respectively. Following filtering, the count data was normalized and log transformed. Highly variable gene selection was performed by `sc.pp.highly_variable_genes(adata, min_mean=0.0125, max_mean=3, min_disp=0.5)`. Highly variable genes were used for subsequent PCA and UMAP projection.

For individual treatment effect analysis, we additionally filter T cell receptor genes, histocompatibility genes, and immunoglobulin genes from the highly variable gene set. Genes to be filtered were obtained from the HUGO database [57]. After filtering, highly variable genes were selected for downstream visualization analysis.

CINEMA-OT analysis was run on each of the samples separately, with `thres=0.5`, `smoothness=1e-4`, `eps=1e-2`, and preweights given by cell types.

For the synergy analysis of donor 3 on day 2 (H3D2), we selected significant synergy genes by a absolute value threshold of 0.15.

References

[1] Maya E Kotas and Ruslan Medzhitov. Homeostasis, inflammation, and disease susceptibility. *Cell*, 160(5):816–827, 2015.

[2] Asif Adil, Vijay Kumar, Arif Tasleem Jan, and Mohammed Asger. Single-cell transcriptomics: current methods and challenges in data acquisition and analysis. *Frontiers in Neuroscience*, 15: 591122, 2021.

[3] Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

[4] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-

seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[5] Lin Yang, Yuqing Zhu, Hua Yu, Xiaolong Cheng, Sitong Chen, Yulan Chu, He Huang, Jin Zhang, and Wei Li. scmageck links genotypes with multiple phenotypes in single-cell crispr screens. *Genome biology*, 21(1):1–14, 2020.

[6] Bin Duan, Chi Zhou, Chengyu Zhu, Yifei Yu, Gaoyang Li, Shihua Zhang, Chao Zhang, Xiangyun Ye, Hanhui Ma, Shen Qu, et al. Model-based understanding of single-cell crispr screening. *Nature communications*, 10(1):1–11, 2019.

[7] Xin Jin, Sean K Simmons, Amy Guo, Ashwin S Shetty, Michelle Ko, Lan Nguyen, Vahbiz Jokhi, Elise Robinson, Paul Oyler, Nathan Curry, et al. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520):eaaz6063, 2020.

[8] Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473): 45–51, 2020.

[9] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[10] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

[11] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

[12] Yunshun Chen, Aaron TL Lun, and Gordon K Smyth. From reads to genes to pathways: differential expression analysis of rna-seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Research*, 5, 2016.

[13] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[14] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022.

[15] Jun Zhao, Ariel Jaffe, Henry Li, Ofir Lindenbaum, Esen Sefik, Ruaidhrí Jackson, Xiuyuan Cheng, Richard Flavell, and Yuval Kluger. Detection of differentially abundant cell subpopulations discriminates biological states in scrna-seq data. *bioRxiv*, page 711929, 2020.

[16] Maren Buettner, Johannes Ostner, Christian L Mueller, Fabian J Theis, and Benjamin Schubert. sccoda is a bayesian model for compositional single-cell data analysis. *Nature communications*, 12(1):1–10, 2021.

[17] Daniel B Burkhardt, Jay S Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature biotechnology*, 39(5):619–629, 2021.

[18] Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

[19] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

[20] Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021.

[21] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[22] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[23] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1):196, 2013.

[24] Sona Vodenkova, Tomas Buchler, Klara Cervena, Veronika Veskrnova, Pavel Vodicka, and Veronika Vymetalkova. 5-fluorouracil and other fluoropyrimidines in colorectal cancer: Past, present and future. *Pharmacology & therapeutics*, 206:107447, 2020.

[25] Lenka Kubiczkova, Lenka Sedlarikova, Roman Hajek, and Sabina Sevcikova. Tgf-β–an excellent servant but a bad master. *Journal of translational medicine*, 10(1):1–24, 2012.

[26] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[27] Chandler Squires and Caroline Uhler. Causal structure learning: a combinatorial perspective. *arXiv preprint arXiv:2206.01152*, 2022.

[28] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[29] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

[30] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[31] Florian Gunsilius and Yuliang Xu. Matching for causal effects via multimarginal optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

[32] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[33] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):1–15, 2017.

[34] Dylan Kotliar, Adrian Veres, M Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A Melton, and Pardis C Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, 8, 2019.

[35] Tomas Eckschlager, Johana Plch, Marie Stiborova, and Jan Hrabeta. Histone deacetylase inhibitors as anticancer drugs. *International journal of molecular sciences*, 18(7):1414, 2017.

[36] Shijie C Zheng, Genevieve Stein-OBrien, Jonathan J Augustin, Jared Slosberg, Giovanni A Carosso, Briana Winer, Gloria Shin, Hans T Bjornsson, Loyal A Goff, and Kasper D Hansen. Universal prediction of cell-cycle position using transfer learning. *Genome biology*, 23(1): 1–27, 2022.

[37] Nagarjuna R Cheemarla, Timothy A Watkins, Valia T Mihaylova, Bao Wang, Dejian Zhao, Guilin Wang, Marie L Landry, and Ellen F Foxman. Dynamic innate immune response determines susceptibility to sars-cov-2 infection and early replication kinetics. *Journal of Experimental Medicine*, 218(8), 2021.

[38] Neal G Ravindra, Mia Madel Alfajaro, Victor Gasque, Nicholas C Huston, Han Wan, Klara Szigeti-Buck, Yuki Yasumoto, Allison M Greaney, Victoria Habet, Ryan D Chow, et al. Single-cell longitudinal analysis of sars-cov-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS biology*, 19(3):e3001143, 2021.

[39] Judith Verhelst, Eef Parthoens, Bert Schepens, Walter Fiers, and Xavier Saelens. Interferon-inducible protein mx1 inhibits influenza virus by interfering with functional viral ribonucleoprotein complex assembly. *Journal of virology*, 86(24):13445–13455, 2012.

[40] Laura Martin-Sancho, Mary K Lewinski, Lars Pache, Charlotte A Stoneham, Xin Yin, Mark E Becker, Dexter Pratt, Christopher Churas, Sara B Rosenthal, Sophie Liu, et al. Functional landscape of sars-cov-2 cellular restriction. *Molecular cell*, 81(12):2656–2668, 2021.

[41] Ilona Jaspers, Katherine M Horvath, Wenli Zhang, Luisa E Brighton, Johnny L Carson, and Terry L Noah. Reduced expression of irf7 in nasal epithelial cells from smokers after infection with influenza. *American journal of respiratory cell and molecular biology*, 43(3):368–375, 2010.

[42] Wenxin Wu, Wei Zhang, J Leland Booth, David C Hutchings, Xiaoqiu Wang, Vicky L White, Houssein Youness, Cory D Cross, Ming-Hui Zou, Dennis Burian, et al. Human primary airway epithelial cells isolated from active smokers have epigenetically impaired antiviral responses. *Respiratory research*, 17(1):1–11, 2016.

[43] Valia T Mihaylova, Yong Kong, Olga Fedorova, Lokesh Sharma, Charles S Dela Cruz, Anna Marie Pyle, Akiko Iwasaki, and Ellen F Foxman. Regional differences in airway epithelial cells reveal tradeoff between defense against oxidative stress and defense against rhinovirus. *Cell reports*, 24(11):3000–3007, 2018.

[44] Magdalena H Hudy, Suzanne L Traves, Shahina Wiehler, and David Proud. Cigarette smoke modulates rhinovirus-induced airway epithelial cell chemokine production. *European Respiratory Journal*, 35(6):1256–1263, 2010.

[45] Boris Landa, Thomas TCK Zhang, and Yuval Kluger. Biwhitening reveals the rank of a count matrix. *arXiv preprint arXiv:2103.13840*, 2021.

[46] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[47] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

[49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[50] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[51] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37 (12):1482–1492, 2019.

[52] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[53] 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, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1):W90–W97, 2016.

[54] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

[55] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

[56] Lindsey W Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M Klein, and Aron B Jaffe. A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte. *Nature*, 560(7718):377–381, 2018.

[57] Susan Tweedie, Bryony Braschi, Kristian Gray, Tamsin EM Jones, Ruth L Seal, Bethan Yates, and Elspeth A Bruford. Genenames. org: the hgnc and vgnc resources in 2021. *Nucleic acids research*, 49(D1):D939–D946, 2021.