

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Mapping coordination and stochasticity of gene regulatory networks at the single cell level

**Permalink**

<https://escholarship.org/uc/item/8mr6k66c>

**Author**

Kim, Min Cheol

**Publication Date**

2024

Peer reviewed|Thesis/dissertation

Mapping coordination and stochasticity of gene regulatory networks at the single cell level

by  
Min Cheol Kim


DISSERTATION  
Submitted in partial satisfaction of the requirements for degree of  
DOCTOR OF PHILOSOPHY

in  
Bioengineering

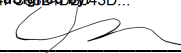
in the  
GRADUATE DIVISION

of the  
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO  
AND  
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:  
  
755C0380968F429... Chun Jimmie Ye  
Chair

DocuSigned by:  
  
Averil Ma

DocuSigned by:  
  
B656F2E231134E5... Patrick Hsu

---

---

Committee Members



## Acknowledgments

This thesis would not have been possible without the neverending support from my family, friends, and the UCSF community. First, I'd like to thank my advisor, Dr. Jimmie Ye, for being my greatest advocate and mentor since the first day of my time at UCSF. I'd also like to thank my thesis committee members Dr. Averil Ma and Dr. Patrick Hsu, who have been crucial to my growth as a scientist. I would also like to thank my parents In Tae Kim and Helen Pak, without whose dedication and support I would not be here today. Finally, I'd like to thank Vivian Wang for providing continuing support and encouragement throughout my career.

Much of the text in this thesis is a reprint of the material as it appears as a preprint in Biorxiv, "memento: Generalized differential expression analysis of single-cell RNA-seq with method of moments estimation and efficient resampling" (doi: [doi.org/mmr8](https://doi.org/10.1101/2018.08.14.244888)). The coauthors listed in this publication directed and supervised the research that forms the basis for the thesis.

# Mapping coordination and stochasticity of gene regulatory networks at the single cell level

Min Cheol Kim

## Abstract

Differential expression analysis of scRNA-seq data is central for characterizing how experimental factors affect the distribution of gene expression. However, it remains challenging to distinguish between biological and technical sources of cell-cell variability and to assess the statistical significance of quantitative comparisons between cell groups. In this thesis, we introduce the statistical method **memento** to address these limitations and enable statistically robust and computationally efficient differential expression analysis of the mean, variability, and gene correlation from scRNA-seq. We used **memento** to analyze 70,000 tracheal epithelial cells to identify interferon response genes with distinct variability and correlation patterns, 160,000 T cells perturbed with CRISPR-Cas9 to reconstruct gene-regulatory networks that control T cell activation, 1.2 million PMBCs to map cell-type-specific *cis* expression quantitative trait loci (eQTLs), and arbitrary cell groups within the entire 50 million cell CELLxGENE Discover data corpus. In all cases, **memento** identified more significant and reproducible differences in mean expression but also identified differences in variability and gene correlation that suggest distinct transcriptional regulation mechanisms imparted by cytokines, genetic perturbations, and natural genetic variation. These results demonstrate **memento** as a first-in-class method for the quantitative analysis of scRNA-seq data, scalable to millions of cells and thousands of samples.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Accurate estimation and inference of distributional parameters from single cell transcriptomics data</b>	<b>5</b>
2.1	Novel statistical model of single-cell RNA-sequencing . . . . .	5
2.2	Estimating distributional parameters of gene expression from scRNA-seq . .	6
2.3	Hypothesis testing using highly efficient bootstrapping . . . . .	9
2.4	Figures . . . . .	12
2.5	Methods . . . . .	14
2.6	Supplementary figures . . . . .	30
<b>3</b>	<b>Generalized differential expression across natural variation and experimental conditions</b>	<b>41</b>
3.1	Differential variability and gene correlation in response to exogenous interferon	41
3.2	Differential expression analysis of perturbed CD4 <sup>+</sup> T cells maps gene regulatory networks in T cell activation . . . . .	44
3.3	Genetic analysis of population-scale single-cell RNA-sequencing . . . . .	47
3.4	Census-scale differential expression analysis across cell types, individuals, and disease states . . . . .	50
3.5	Figures . . . . .	53

3.6	Methods . . . . .	57
3.7	Supplementary figures . . . . .	65
<b>4</b>	<b>Discussion and conclusion</b>	<b>72</b>
	<b>References</b>	<b>76</b>

## List of Figures

2.1	memento workflow for differential mean, variability, and gene correlation testing . . . . .	12
2.2	Performance of memento in simulation and real data . . . . .	13
3.1	Mapping transcriptional response of human bronchial epithelial cells to extracellular interferon using memento . . . . .	53
3.2	Reconstructing gene regulatory networks of T cell activation using Perturb-seq and memento . . . . .	54
3.3	Mapping of mean QTL, variability QTL, and correlation QTL using memento . . . . .	55
3.4	Extending memento for near real-time differential expression analysis within CZI CELLxGENE Discover . . . . .	56



# Chapter 1

## Introduction

Gene expression, inherently determined by a cell's genetic constitution and its environmental interactions, can exhibit fluctuations due to both intrinsic noise (stemming from mRNA transcription and degradation) and extrinsic noise related to a cell's specific state<sup>1,2</sup>. While genetics and environmental history significantly contribute to expression variability across a population of cells, stochastic transcriptional noise has been recognized to profoundly influence cellular responses to perturbations, as well as cellular development and differentiation<sup>2-4</sup>. Characterizing how deterministic and stochastic factors cohesively influence the distribution of gene expression is central for understanding how transcriptional control is established, maintained, and may be broken. These insights could, consequently, spotlight mechanisms underpin phenomena where genotype-phenotype relationships are not completely explained, such as destabilization<sup>3</sup>, incomplete penetrance<sup>5</sup>, and variable expressivity<sup>6</sup>.

Historically, the distribution of a gene's expression across a cellular population has been described by parameters such as the mean and variance, as well as derivative measures like the Fano factor and coefficient of variation<sup>7</sup>. Constitutively expressed genes, such as housekeeping genes, which undergo transcription and degradation at constant rates, are

predicted to conform to a Poisson distribution where the mean and variance are equal. Nonetheless, most genes display over-dispersion, exhibiting higher variability than expected<sup>8</sup>, and genes within the same biological pathway are often transcriptionally correlated<sup>5</sup>. These observations are consistent with a model of active regulation for multiple related genes controlled by *cis* regulatory elements for transcription factors with "on" and "off" states<sup>9</sup>. Until recently, studying the distribution of gene expression, in particular the joint distribution of multiple genes, has been technologically challenging and has been mostly pursued in model organisms that can be genetically modified<sup>10,11</sup>.

Single-cell RNA-sequencing (scRNA-seq) has emerged as a systematic and efficient approach for profiling the transcriptomes of cells across experimental factors including extracellular stimuli<sup>12</sup>, genetic perturbations<sup>13,14</sup>, and natural genetic variation<sup>15-18</sup>. In theory, the analysis of scRNA-seq data can decipher how experimental factors and transcriptional noise together shape the distribution of gene expression. Yet, there remains a need for differential expression analysis methods that compare distributional parameters between cell groups including the mean, variability, and gene correlation. To assess differences in mean expression, it is common practice to apply bulk RNA-seq differential expression analysis methods to pseudobulk profiles, generated by aggregating transcript counts for cell groups defined by clustering. While not fully leveraging single cells as repeated measures, pseudobulk approaches have surprisingly proven to outperform methods that explicitly model the distribution of observed scRNA-seq data<sup>19</sup>. Moreover, very few methods exist for assessing differences in gene expression variability that measure transcriptional noise or correlation between pairs of genes that measure the coordinated expression of genes that may participate in the same regulatory network.

Generalized differential expression analysis of scRNA-seq data remains a formidable challenge due to two pivotal statistical limitations. First, decomposing the overall cell-to-cell variability in scRNA-seq data into its constituent components - transcriptional and

measurement noise - remains a substantive obstacle<sup>20</sup>. This complexity arises due to the small numbers of molecules required in the biochemical reactions of both gene transcription and the scRNA-seq sampling process (**Fig. 2.1A**)<sup>21</sup>. Most existing methods implement highly parameterized models designed to account for the higher than expected variability in the *observed* sparse transcript counts, attributing it to known experimental and technical factors (i.e. treatment and batch). However, these models do not explicitly model measurement noise, a byproduct of the inherent undersampling characteristics of scRNA-seq workflows<sup>22-27</sup>. Importantly, generating accurate estimates of biological variability is crucial for effectively modeling the joint distribution of multiple genes, which is often represented by the correlation between gene pairs<sup>22</sup>. Second, establishing the statistical significance of a specific comparison of mean, variability, or gene correlation between groups of cells, remains a largely unsolved problem. Many existing methods utilize asymptotic theory for comparison of means, leading to uncalibrated p-values, demand an exact specification of the parametric model, or lack the flexibility to incorporate hierarchical structures and continuous covariates effectively. This is particularly problematic in studies necessitating thousands of comparisons, as inadequately calibrated p-values violate assumptions for multiple testing correction. Moreover, while multiplexed workflows inherently accommodate a growing number of individuals or conditions, thereby naturally generating biological and technical replicates<sup>13,15,28-30</sup>, the majority of existing analytical methods do not explicitly account for such replicates in their models. DESCEND and comparable methods that utilize generalized linear models are equipped to address this issue. However, they often encounter significant computational hurdles when modeling the complex hierarchical structure inherent in scRNA-seq data. Furthermore, these methods are limited to a specific model of cell-cell variability<sup>31</sup>, such as the proportion of cells exhibiting zero expression. Indeed, recent studies spotlight a startling underperformance of scRNA-seq methods relative to pseudobulk methods when testing mean differences, likely attributable to limitations in both multiple testing correction and properly accounting for the hierarchical nature of scRNA-seq data<sup>19</sup>.

Addressing these statistical and methodological challenges, we present **memento**, an end-to-end method that implements a hierarchical model for estimating mean, residual variance, and gene correlation from scRNA-seq data, and provides a statistical framework for hypothesis testing of these parameters (**Fig. 2.1B**). **memento** employs a novel multivariate hypergeometric sampling process and leverages the sparsity of scRNA-seq data to implement an innovative bootstrapping strategy for the efficient statistical comparison of the estimated parameters between cell groups. Through simulations and analyses of real data, we demonstrate that **memento** produces accurate parameter estimates over a range of gene expression distributions and sampling efficiencies, computes well-calibrated test statistics suitable for multiple testing correction, and achieves sublinear runtimes. We demonstrate the broad applicability of **memento** in four applications aimed at elucidating how experimental factors modulate the distribution of gene expression in human cells (**Fig. 2.1C**). First, we conducted scRNA-seq on 70k tracheal epithelial cells stimulated with extracellular interferons and investigated how stimulation modulates the variability and correlation of response genes temporally. Second, we performed Perturb-seq on 160k T cells and mapped gene regulatory networks that define aspects of broad T cell activation. Third, we reanalyzed 1.2M cells collected from 250 individuals to identify genetic variants associated with mean, variability, and gene correlation in specific cell types. Finally, we implemented an approximate bootstrapping strategy utilizing the Chan Zuckerberg Initiative (CZI) CELLxGENE Discover Census API, facilitating the deployment of **memento** for near real-time comparisons of any arbitrary cell groups within the 50 million cell CELLxGENE data corpus. Across these diverse applications, **memento** consistently identified more significant and reproducible differences in mean expression between experimental groups compared to existing methods but also identified differences in variability and gene correlation, thereby revealing distinct transcriptional regulation modes imparted by cytokines, genetic perturbations, and natural genetic variation. **memento** is implemented in python, is compatible with scanpy<sup>32</sup>, and can be downloaded at <https://github.com/yelabucsf/scrna-parameter-estimation>.

# Chapter 2

## Accurate estimation and inference of distributional parameters from single cell transcriptomics data

### 2.1 Novel statistical model of single-cell RNA-sequencing

Since its advent, single-cell RNA-sequencing (scRNA-seq) has yielded sparse data despite continuous advancements in molecular biology, manifesting in a high degree of cell-to-cell variability even in genetically identical cells exposed to the same environment (**Fig. 2.1A**). The crucial task of decomposing this variability into components of biological versus measurement noise becomes pivotal for the differential expression analysis of scRNA-seq data. Notably, measurement noise intrinsic to scRNA-seq can be attributed to inefficiencies in at least three molecular biology processes common to nearly all workflows: 1) the capture of only a fraction of expressed transcripts within compartments for reverse transcription (RT) to cDNA, 2) the amplification of only a fraction of cDNA molecules during each polymerase chain reaction (PCR) cycle, and 3) the sequencing of only a fraction of the amplified cDNA. Although the development of Unique Molecular Identifiers (UMIs) has largely obviated the need to model the noise introduced by PCR<sup>33</sup>, noise stemming from imperfect transcript capture for RT and imperfect cDNA sampling during sequencing

persists, culminating in the observed, attenuated distribution of counts.

Here, we propose a novel statistical framework that models observed scRNA-seq counts as the result of hypergeometric sampling of the expressed transcripts within a cell. The motivation to implement the hypergeometric model stems from the observation that the capture of poly-adenylated mRNA for RT and sequencing of resultant libraries are processes which sample molecules from each cell without replacement, thereby introducing measurement noise into the final dataset. Central to our model is the flexibility to accommodate arbitrary distributions of gene expression within a cell prior to measurement. Formally, let  $\mathbf{X}_c = \frac{\mathbf{Z}_c}{N_c}$  denote an  $m$ -dimensional random variable representing the normalized transcript counts of  $m$  genes in cell  $c$ , where  $\mathbf{Z}_c$  defines a vector of the expressed transcript counts and  $N_c$  the total transcript counts within a cell. We model scRNA-seq as a multivariate hypergeometric sampling process, wherein the observed transcript counts  $\mathbf{Y}_c$  originate from  $\mathbf{X}_c$ :  $\mathbf{Y}_c \sim \text{MultiHG}(N_c \mathbf{X}_c, N_c, N_c q)$ . In this representation,  $q$  signifies the overall transcript sampling efficiency of scRNA-seq and is associated with measurement noise introduced during library preparation and sequencing (see **Methods** for detailed exploration). Importantly, we empirically substantiate that the two-step noise process involving RT capture (hypergeometric) and sequencing (binomial) can be well represented with a singular step of hypergeometric sampling with the overall  $q$  (**Fig. S2.1**). Across many simulated values of capture efficiency and sequencing saturation, the single hypergeometric sampling well approximates the two step process (nonsignificant KS-test, **Fig. S2.2**).

## 2.2 Estimating distributional parameters of gene expression from scRNA-seq

To our knowledge, this is the first use of the hypergeometric sampling process for modeling scRNA-seq data, a likely result of the complexity in estimating distribution parameters via

maximum likelihood. Here, we derive method of moment (MoM) estimators for the first (mean), second (variance), and mixed (covariance) moments of  $\mathbf{X}_c$  given  $\mathbf{Y}_c$  under the assumption of hypergeometric sampling (see **Methods** for derivation and details):

$$\begin{aligned}\hat{\mu}_{g,\text{memento}} &= \frac{1}{\sum_c N_c} r_g^*, \text{ where } r_g^* \text{ is the Good-Turing corrected count of gene } g \\ \hat{\sigma}_{g,\text{memento}}^2 &= \frac{1}{n_{\text{cells}}} \sum_c \frac{Y_{cg}^2 - Y_{cg}(1-q)}{N_c^2 q^2} - \hat{\mu}_g^2 \\ \hat{\sigma}_{g_i g_j, \text{memento}} &= \frac{1}{n_{\text{cells}}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2} - \hat{\mu}_{g_i} \hat{\mu}_{g_j}\end{aligned}$$

While the mean can be directly used to test for differential mean expression (DM), the variance needs to be adjusted to account for the expected dependence between mean and variance in count data, thereby enabling the testing for differential expression variability (DV) independent of DM<sup>34,35</sup>. To do so, we introduce the *residual variance*  $\tilde{\sigma}_g$  as a measure of expression variability  $\sigma_g$  (**Methods**), defined as the variance component unexplained by the mean (**Methods**). Consequently, gene correlations are the covariance terms (off diagonal elements) scaled by the variance terms (diagonal elements) from the variance-covariance matrix estimated above.

We performed extensive simulations to compare **memento**'s hypergeometric estimators to the naive plug-in estimators employed by scHOT<sup>36</sup>, empirical Bayes estimators under the Poisson approximation introduced by Zhang et al.<sup>37</sup> (a special case of the **memento** estimator for setting  $q = 0$ , and estimates derived from BASiCS<sup>27</sup> (see **Methods** for forms of the naive and Poisson estimators). Across a range of  $q$  values, reported by both low-efficiency ( $q < 0.2$ ) droplet-based (e.g., 10X V1, V2 and V3) and high-efficiency ( $q > 0.3$ ) plate-based (Smart-Seq3<sup>38</sup>) scRNA-seq workflows, **memento**'s hypergeometric estimator produced remarkably accurate estimates of mean (Lin's concordance correlation coefficient -  $\rho_c > 0.98$  with 100 cells,  $> 0.8$  with 10 cells), residual variance ( $\rho_c > 0.98$ , 100 cells), and gene correlation ( $\rho_c > 0.98$ , 100 cells) (**Fig. 2.2A**). In addition, **memento** produces stable residual

variance and gene correlation estimates across  $qs$ , outperforming other estimators for both low- and high-efficiency workflows. To investigate how a gene’s mean expression influence the accuracy of variance estimation, we compared the true simulated mean with the average error in variance estimation (**Fig. S2.3**). We found that all estimators have higher accuracy for highly expressed genes, but **memento** outperforms other methods even for lowly expressed genes. These simulations are based on a single-step sampling approach, which, as demonstrated above, effectively approximates the two-step sampling process modeling RT and sequencing.

To further validate the accuracy of **memento**’s parameter estimates, we reanalyzed a dataset comprising paired droplet-based scRNA-seq and RNA FISH data<sup>39</sup>. This data was previously analyzed using SAVER<sup>40</sup>, an imputation method that borrows information from similar genes and cells that has been shown to outperform other approaches including MAGIC and scImpute for estimating gene correlations (**Fig. 2.2B**). For genes profiled using both scRNA-seq and FISH, **memento**’s mean estimates exhibited modest improvements over the naive estimator used by SAVER, scHOT, and BASiCs (21 genes considered;  $\rho = 0.58$  and  $\rho = 0.54$ , using 100 cells). For residual variance, **memento**’s estimates were significantly more correlated with those obtained by FISH (14 genes considered;  $\rho = 0.71$ ) than the naive estimator ( $\rho = 0.56$ ) and BASiCS ( $\rho = 0.61$ ) using all available 8,498 cells. Finally for gene correlation, **memento** ( $\rho = 0.53$ ) also significantly outperforms the naive estimator ( $\rho = 0.29$ ), SAVER ( $\rho = 38$ ), and scVI ( $\rho = 23$ ) using all cells. Importantly, **memento** produces better estimates of gene correlation without utilizing additional genes required by imputation methods (e.g., SAVER) and variational inference methods (e.g., SCVI). This advantage translates not only to computational efficiency in estimation (**memento**: 17 seconds vs SAVER: 30 minutes for 14 gene pairs) but also produces estimates that might be better suited for specific downstream analyses, such as genetic mapping, where imputation could inadvertently introduce confounding effects by borrowing information across genes and cells. These results underscore the accuracy of **memento**’s parameter estimates, demonstrated



through both simulations and comparative analyses against benchmark FISH data.

## 2.3 Hypothesis testing using highly efficient bootstrapping

The goal for hypothesis testing is to determine if an observed difference in estimated parameters between cell groups, such as mean, variability, and gene correlation, is statistically significant in comparison to a null hypothesis. Notably, the primary concern when testing thousands of genes, typical in scRNA-seq experiments profiling the entire transcriptome, is the multiple testing problem: nominating a feasible set of candidate genes for experimental follow-up while predicting the expected number of validations.

Consequently, apt calibration of the distribution of test statistics under the null hypothesis, amenable to multiple testing correction, becomes imperative. Although method of moments estimates utilizing the hypergeometric model offers simplicity in computation and flexibility to the true gene expression distribution within cells, establishing statistical significance and computing confidence intervals (CIs) necessitate bootstrapping the data. Bootstrapping large numbers of cells using a standard scheme that samples cells with replacement would require extensive computational resources that are both time and memory prohibitive, especially for large datasets.

**memento** implements an innovative scheme, capitalizing on the sparsity and discreteness of scRNA-seq data, to facilitate fast, memory-efficient, and highly parallelizable bootstrapping. The key to our scheme resides in recognizing that the number of unique observed transcript counts is substantially smaller than the number of cells (**Fig. S2.5**), and this held true even for unique observed pairs of counts (**Fig. S2.6**) albeit to a lesser extent. Therefore, each bootstrap iteration necessitates merely the resampling of  $K$  unique transcript counts for each gene from  $\text{Multinomial}(N, \frac{n_1}{N} \dots \frac{n_K}{N})$ , proportional to the observed frequency of each count

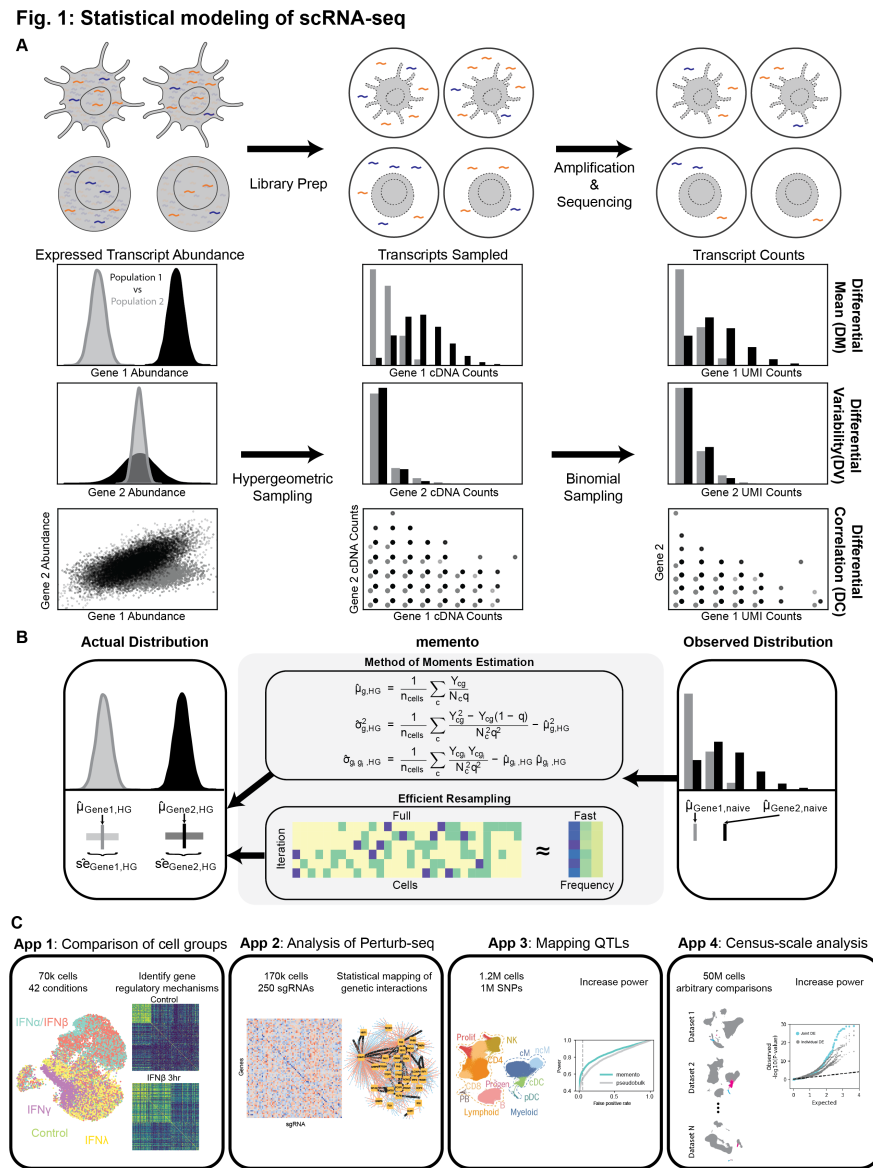
(**Fig. S2.4**), as opposed to resampling individual cells' counts from a multinomial distribution comprising  $N$  elements (cells) ( $\text{Multinomial}(N, \frac{1}{N} \dots \frac{1}{N})^{41}$ ). This approach culminates in fitting a markedly small weighted dataset ( $K \ll N$ ) for each resampling iteration. To accommodate multiplexed experiments, we extend our bootstrapping strategy using a meta-regression framework, considering each replicate as a separate subgroup of the data, thereby enabling hierarchical resampling. This approach allows us to quantify uncertainty while respecting the process with which the data was generated, such as sampling of cells from different individuals. In simulation, **memento**'s bootstrapping strategy yields highly accurate estimates of the null distribution for mean, residual variance, and gene correlation comparable to those obtained with naive bootstrap resampling across a wide range of genes (**Fig. S2.8, Fig. S2.7**). Utilizing bootstrap to quantify the CI in parameter estimates, **memento** computes well-calibrated empirical p-values for DM, DV, and DC, suitable for multiple testing correction (**Fig. S2.9**).

To show that **memento** produces well-calibrated p-values while maintaining high statistical power, we simulated a dataset encompassing two distinct cell populations. To maintain relevance to actual data, parameters extracted from a real dataset of helper T cells pre and post-stimulation with rIFN $\beta$  were employed. In the simulation, differences in the mean, variability, and correlation were retained for 150 genes and removed for the remainder (see **Methods**). For the differential mean simulation, we created the dataset with biological replicates to mimic multiplexed experimental designs<sup>17</sup>. We show that for DM, DV, and DC, **memento** computed well calibrated p-values with the expected number of false positives at a specified significance cutoff, while achieving the highest power for detecting true differences (**Fig. 2.2C**). Especially for DV and DC tasks, **memento** vastly outperformed competing methods in power, maintaining a reduced false positive rate at each significance threshold. Moreover, we observed that established methods for DM are either too liberal (t-test, Wilcoxon rank-sum test) or far too conservative (edgeR, DESeq2), consistent with results from<sup>19</sup> (**Fig. S2.10**). Squair et al. previously attributed this to replicate-level heterogeneity

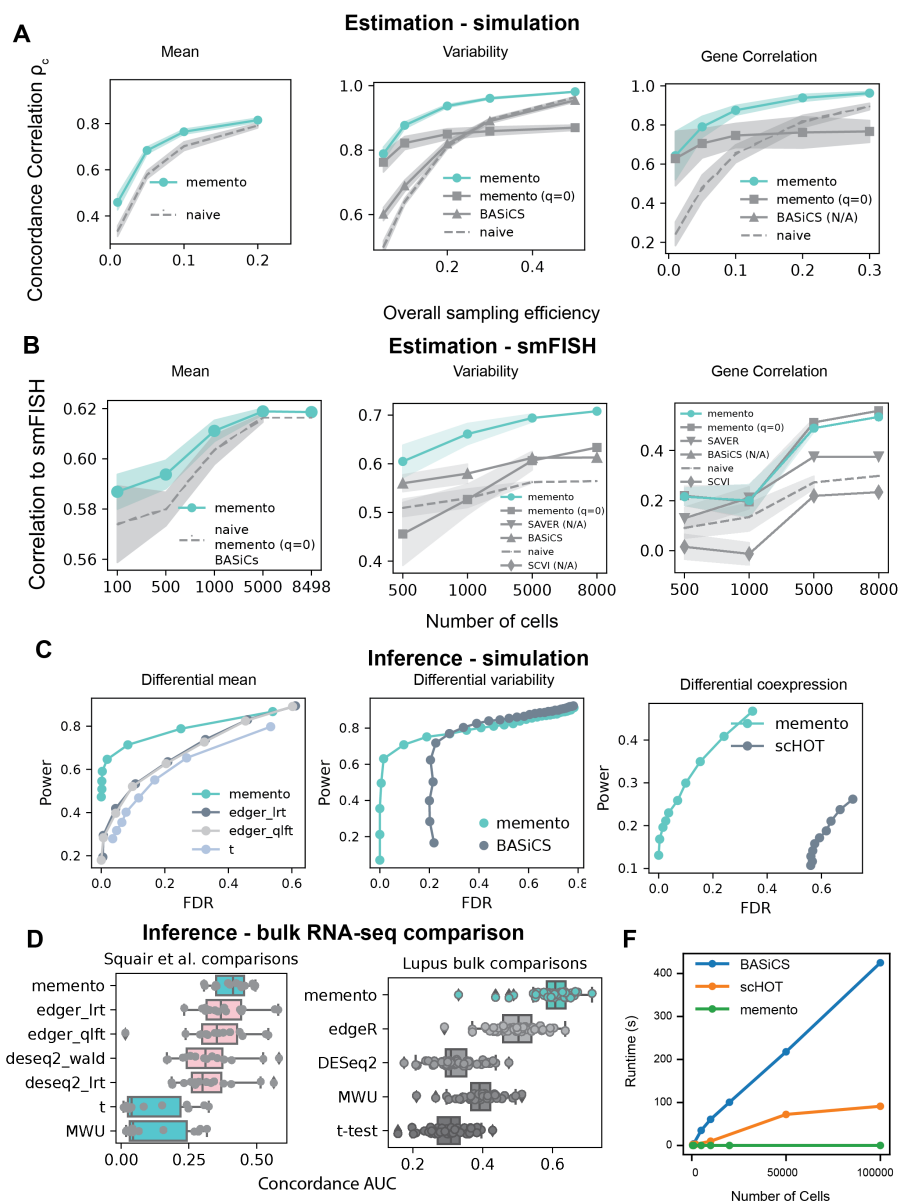
present in most scRNA-seq datasets, and recommended pseudobulk methods to simplify the hierarchical structure<sup>19</sup>. By directly accounting for this structure, **memento** produced expected false positive rates at each significance threshold even when varying degrees of heterogenous effects are present. In addition to simulations, we also benchmarked **memento** using paired single-cell and bulk RNA-seq samples, employing datasets used by Squair et al.<sup>19</sup> (**Fig. 2.2D left**) and an additional dataset from lupus patients (**Fig. 2.2D right**)<sup>42</sup>. In both datasets, **memento** was able to produce DM results from the scRNA-seq data most concordant with those obtained from analyses of bulk RNA-seq. Finally, we showed that **memento** finds the most concordant DM genes when comparing the effects of IFN- $\alpha$  and IFN- $\beta$  on ciliated cells, both of which stimulated the identical type-1 interferon receptor (**Fig. S2.11**).

Compared to existing methods for DM, DV, and DC, **memento** is able to perform hypothesis testing at computational speeds that are orders of magnitude faster, enabling scalability to millions of cells (**Fig. 2.2F**). In a scenario simulating throughput analagous to emerging scRNA-seq datasets - two groups each containing  $10^6$  cells - conducting DM and DV analysis for 1,000 genes using 10,000 bootstrapping iterations per gene between groups necessitated a mere 13 minutes using a single CPU. A multicore implementation of **memento** facilitated the parallelization of multiple genes, curtailing runtime to 2-3 minutes with 6 CPUs. Particularly for DV and DC, **memento** achieves up to 1000x gain in computational speed given the equivalent compute resources compared to existing methods. These results substantiate that **memento**'s bootstrapping strategy yields accurate CI estimates for effect sizes at high computational efficiency, culminating in well-calibrated test statistics and enabling hypothesis testing of scRNA-seq data scalable to groups of millions of cells (see **Methods** for detailed description of the resampling strategy and hypothesis testing).

## 2.4 Figures



**Figure 2.1: memento workflow for differential mean, variability, and gene correlation testing (A)** Experimental workflow for single cell RNA-sequencing samples RNA transcripts inside each cell during library preparation and sequencing. **(B)** memento models scRNA-seq as a hypergeometric sampling process, estimates expression distribution parameters (mean, residual variance, and correlation) using method of moments estimators, implements bootstrapping for estimating confidence intervals, and tests for differences in expression parameters between two groups of cells. **(C)** Four applications of memento.



**Figure 2.2: Performance of memento in simulation and real data (A)** Concordance of estimates of mean (left), variability (middle), and gene correlation (right) with simulated ground truth values (y-axis) for a range of overall transcript capture efficiencies ( $q$ ) (x-axis). **(B)** Correlation of estimates of mean (left), variability (middle), and gene correlation (right) of DropSeq data against smFISH-derived estimates measured in the same population of melanoma cells across different numbers of DropSeq cells used. **(C)** Power (y-axis) vs FDR (x-axis) comparing existing methods with memento for DM (left), DV (middle), and DC (right). **(D)** Concordance AUC (x-axis) of single-cell differential mean analysis to pseudobulk differential mean analysis using datasets in<sup>19</sup> (left) and<sup>17</sup>. **(E)** Runtime (y-axis) of three methods across number of cells (x-axis) for DM/DV.

## 2.5 Methods

### Modeling scRNA-seq as a hypergeometric sampling process

We model the count data obtained from scRNA-seq with a flexible hierarchical model that explicitly considers the generative process of the expressed transcript counts and sampling of mRNA molecules with massively-parallel scRNA-seq methods. As presented in the main text, our full model of the scRNA-seq sampling process can be summarized as follows:

$$\begin{aligned} \mathbf{Z}_c &\sim P_Z, \text{ expressed transcript counts in cell } c \\ N_c &= \mathbf{1}^T \mathbf{Z}_c, \text{ total transcript count of cell } c \\ \mathbf{X}_c &= \frac{\mathbf{Z}_c}{N_c}, \text{ normalized transcript counts in cell } c \end{aligned}$$

$$\mathbf{Y}_c \sim \text{MultiHG}(\mathbf{Z}_c, N_c, qN_c) = \text{MultiHG}(\mathbf{X}_c N_c, N_c, qN_c), \text{ observed transcript counts in cell } c$$

$q$  is the random variable representing the proportion of expressed transcript counts that is eventually counted as UMIs in the observed scRNA-seq experiment. In our discussion of sources of noise above as applied to most scRNA-seq workflows, it accounts for both the RT sampling efficiency as well as the sampling of transcripts from sequencing. In the extreme, if a library is sequenced to saturation, then  $q$  reduces to the RT sampling efficiency; however, in most experiments, libraries are not sequenced to saturation but up to a known percentage of unique molecules. Through extensive simulations, we demonstrate that this compound noise process can be well approximated with a single multivariate hypergeometric process by using a value for  $\mathbb{E}[q]$  that is a product of the RT sampling efficiency (available for specific experimental technologies) and the sequencing sampling efficiency (available from the preprocessing pipelines such as CellRanger) (**Fig. S2.1**)<sup>43</sup>.

We then model the mRNA capture process with a multivariate hypergeometric distribution. The probability mass function (PMF) of the multivariate hypergeometric distribution given  $(Z_1, Z_2, Z_3, \dots, Z_G)$  components (i.e. genes), total count  $N = \sum_{i=1}^G Z_i$ , and number of samples  $n \in 0, 1, \dots, N$  is given by:

$$p_{\text{MultiHG}}(\mathbf{Y}; Z_1, Z_2, \dots, Z_G, N, n) = \frac{\prod_{i=1}^G \binom{Z_i}{Y_i}}{\binom{N}{n}} \quad (2.1)$$

In previous works<sup>37</sup>, the full hypergeometric treatment was simplified by a series of approximations, starting from the hypergeometric model to the Poisson model:

$$\begin{aligned} \mathbf{Y}_c &\sim \text{MultiHG}(\mathbf{Z}_c, N_c, qN_c), \text{ observed transcript counts} \\ \mathbf{Y}_c &\sim \text{Multinomial}\left(\frac{\mathbf{Z}_c}{N_c}, qN_c\right), \text{ observed transcript counts} \\ Y_{cg} &\sim \text{Bn}\left(\frac{Z_{cg}}{N_c}, qN_c\right), \text{ observed transcript counts} \\ Y_{cg} &\sim \text{Poi}\left(\frac{Z_{cg}}{N_c} qN_c\right), \text{ observed transcript counts} \\ Y_{cg} &\sim \text{Poi}(qZ_{cg}), \text{ observed transcript counts} \end{aligned}$$

$Y_{cg}$  is a single element in the vector  $\mathbf{Y}_c$ , as the Poisson model considers the sampling of each gene to be independent. As we discuss in the following sections, the full hypergeometric treatment and the Poisson simplification result in very similar estimators when  $q$  is very small (close to 0), but become more different as the value of  $q$  increases, as scRNA-seq experimental workflow improves.

## Method of moments estimation of expressed transcript counts

We will start this section by reviewing the derivation of the Poisson estimators first presented in<sup>37</sup> in the context of determining optimal sequencing depth for scRNA-seq experiments. First, recall the previously presented Poisson sampling model for scRNA-seq where  $N_c$  represents the total expressed transcripts for each cell,  $q$  is the overall sampling efficiency, and  $X_{cg}$  is the true relative mRNA expression  $Y_{cg} \sim \text{Poi}(qN_cX_{cg})$ .

For a Poisson variable  $A \sim \text{Poi}(\lambda)$ , the moments of  $A$  are  $\mathbb{E}[A] = \lambda$  and  $\mathbb{E}[A^2] = \lambda^2 + \lambda$ .

Similarly, for our model, we can write down the equations for the moments of  $Y_{cg}$  given the other variables,  $q$ ,  $N_c$ , and  $X_c$ .

$$\begin{aligned}\mathbb{E}[Y_{cg}|X_{cg}, N, q] &= X_{cg}Ncq \\ \mathbb{E}[Y_{cg}^2|X_{cg}, N, q] &= X_{cg}^2N_c^2q_c^2 + X_{cg}q_cN_c \\ \mathbb{E}[Y_{cg_i}Y_{cg_j}|X_{cg_i}, X_{cg_j}, N, q] &= \mathbb{E}[X_{cg_i}X_{cg_j}N_c^2q^2|X_{cg_i}, X_{cg_j}, N, q] = X_{cg_i}X_{cg_j}N_c^2q^2\end{aligned}\tag{2.2}$$

Substituting the first moment equation into the second, we get:

$$\mathbb{E}[Y_{cg}^2 - Y_{cg}|X_{cg}, N, q] = X_{cg}^2N_c^2q^2\tag{2.3}$$

These equations lead to an estimator for  $\hat{\mu}_{g,Poi}$ ,  $\hat{\sigma}_{g,Poi}^2$ , and  $\hat{\sigma}_{g_i g_j, Poi}$ , the mean, variance, and covariance of  $X_{cg}$  by averaging the moments over all cells:



$$\begin{aligned}
\hat{\mu}_{g,Poi} &= \hat{\mathbb{E}}[X_{cg}] = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \\
\hat{\sigma}_{g,Poi}^2 &= \hat{\mathbb{E}}[X_{cg}^2] - \hat{\mathbb{E}}[X_{cg}]^2 = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}^2 - Y_{cg}}{N_c^2 q^2} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \right)^2 \\
\hat{\sigma}_{g_i g_j, Poi} &= \hat{\mathbb{E}}[X_{cg_i} X_{cg_j}] - \hat{\mathbb{E}}[X_{cg_i}] \hat{\mathbb{E}}[X_{cg_j}] = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2} \\
&\quad - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i}}{N_c q} \right) \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_j}}{N_c q} \right)
\end{aligned} \tag{2.4}$$

Now, let us consider the full multivariate hypergeometric model,

$\mathbf{Y}_c \sim \text{MultiHG}(\mathbf{X}_c N_c, N_c, q N_c)$ . For a random vector  $\mathbf{A} \sim \text{MultiHG}(\mathbf{K}, N, n)$ , the moments of  $\mathbf{A}$  are:

$$\begin{aligned}
\mathbb{E}[A_i] &= n \frac{K_i}{N} \\
\mathbb{E}[A_i^2] &= n \frac{N-n}{N-1} \frac{K_i}{N} \left( 1 - \frac{K_i}{N} \right) + n^2 \frac{K_i^2}{N^2} \\
\mathbb{E}[A_i A_j] &= -n \frac{N-n}{N-1} \frac{K_i K_j}{N^2} + n^2 \frac{K_i K_j}{N^2}
\end{aligned} \tag{2.5}$$

We can again write down the moment equations, this time for the multivariate hypergeometric model.

$$\begin{aligned}
\mathbb{E}[Y_{cg}|X_{cg}, N_c, q] &= qN_c \frac{X_{cg}N_c}{N_c} \\
&= X_{cg}N_cq \\
\mathbb{E}[Y_{cg}^2|X_{cg}, N_c, q] &= qN_c \frac{N_c - qN_c}{N_c - 1} \frac{X_{cg}N_c}{N_c} \left(1 - \frac{X_{cg}N_c}{N_c}\right) + q^2N_c^2 \frac{X_{cg}^2N_c^2}{N_c^2} \\
&\approx qN_c(1 - q)X_{cg}(1 - X_{cg}) + q^2N_c^2X_{cg}^2 \\
&= X_{cg}^2N_c^2q^2 + X_{cg}qN_c(1 - q) - X_{cg}^2qN_c(1 - q) \\
&= X_{cg}^2(N_c^2q^2 - N_cq(1 - q)) + X_{cg}N_cq(1 - q) \\
\mathbb{E}[Y_{cg_i}Y_{cg_j}|X_{cg_i}, X_{cg_j}, N, q] &= -qN_c \frac{N_c - qN_c}{N_c - 1} \frac{X_{cg_i}X_{cg_j}N_c^2}{N_c^2} + q^2N_c^2 \frac{X_{cg_i}X_{cg_j}N_c^2}{N_c^2} \\
&\approx q^2N_c^2X_{cg_i}X_{cg_j} - q(1 - q)N_cX_{cg_i}X_{cg_j} \\
&= X_{cg_i}X_{cg_j}(N_c^2q^2 - N_cq(1 - q))
\end{aligned} \tag{2.6}$$

Substituting the first moment equation into the second, we get:

$$\mathbb{E}[Y_{cg}^2 - (1 - q)Y_{cg}|X_{cg}, N, q_c] = X_{cg}^2(N_c^2q^2 - N_cq(1 - q)) \tag{2.7}$$

The approximation used in the derivation for the second and first pairwise moment assumes that  $N_c \gg 1$ . For most mammalian cells with expressed transcript counts on the order of  $10^5$ , these approximation should hold. Similar to estimators based on the Poisson model, we can derive estimators based on these moment equations from the full multivariate hypergeometric model:

$$\begin{aligned}
\hat{\mu}_{g,HG} &= \hat{\mathbb{E}}[X_{cg}] = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \\
\hat{\sigma}_{g,HG}^2 &= \hat{\mathbb{E}}[X_{cg}^2] - \hat{\mathbb{E}}[X_{cg}]^2 \\
&= \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}^2 - Y_{cg}(1-q)}{N_c^2 q^2 - N_c q(1-q)} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \right)^2 \\
&\approx \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}^2 - Y_{cg}(1-q)}{N_c^2 q^2} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \right)^2 \tag{2.8} \\
\hat{\sigma}_{g_i g_j, HG} &= \hat{\mathbb{E}}[X_{cg_i} X_{cg_j}] - \hat{\mathbb{E}}[X_{cg_i}] \hat{\mathbb{E}}[X_{cg_j}] \\
&= \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2 - N_c q(1-q)} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i}}{N_c q} \right) \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_j}}{N_c q} \right) \\
&\approx \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i}}{N_c q} \right) \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_j}}{N_c q} \right)
\end{aligned}$$

Last, we write the naive estimators for mean, variance and covariance for completeness.

$$\begin{aligned}
\hat{\mu}_{g,naive} &= \hat{\mathbb{E}}[X_{cg}] = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \\
\hat{\sigma}_{g,naive}^2 &= \hat{\mathbb{E}}[X_{cg}^2] - \hat{\mathbb{E}}[X_{cg}]^2 = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}^2}{N_c^2 q^2} - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg}}{N_c q} \right)^2 \\
\hat{\sigma}_{g_i g_j, naive} &= \hat{\mathbb{E}}[X_{cg_i} X_{cg_j}] - \hat{\mathbb{E}}[X_{cg_i}] \hat{\mathbb{E}}[X_{cg_j}] = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2} \\
&\quad - \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i}}{N_c q} \right) \left( \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_j}}{N_c q} \right) \tag{2.9}
\end{aligned}$$

So far, the estimators for the mean and covariance is identical between the naive, Poisson and HG estimators. However, the estimator for the variance, which contributes to the measurement of residual variance and correlation, is the key difference between the three sets of estimators. Importantly, it is straightforward to see that the HG estimator for the variance includes the naive and Poisson estimators:

$$\begin{aligned}\lim_{q \rightarrow 0} \hat{\sigma}_{g,HG}^2 &= \hat{\sigma}_{g,Poi}^2 \\ \lim_{q \rightarrow 1} \hat{\sigma}_{g,HG}^2 &= \hat{\sigma}_{g,naive}^2\end{aligned}\tag{2.10}$$

These results imply that when  $q$ , the overall sampling efficiency, is small, the HG estimators behave very similar to the Poisson estimators. When  $q$  approaches 1, a hypothetical scenario where the scRNA-seq workflow is perfect and we capture all expressed transcripts, the HG estimators converge to the naive estimator, as there is no noise process. As scRNA-seq workflows improve and  $q$  becomes larger, HG estimators serve as a generalization of the estimators presented by Zhang et al. to account for different types of experimental workflows with different values of  $q$ .

We also discuss here the case where  $q$  is not constant across cells. One of the assumptions used in deriving our estimators above is that  $q$  is a known constant, and we do not need to estimate it for each and every cell. However, it is plausible that for certain scRNA-seq technologies and when sequencing is not saturated,  $q$  is actually a distribution around its mean,  $\mathbb{E}[q]$ . Experimentally, we can mitigate this issue by using spike-in RNA control to actually measure the value of  $q$  for each and every cell. Because  $q$  does not appear in the Poisson estimators, it is not possible to explicitly account for the variability in  $q$  even if its value can be measured for each cell. With the hypergeometric estimators derived here, we can simply substitute the measured values of  $q_c$  for each cell in place of  $q$  above.

## Improving mean estimation with Good-Turing method

Our derivations in the section show different naive, Poisson and hypergeometric estimators for variance and correlation, but the mean estimator were identical. We sought to further improve the mean estimation especially for small population of cells, which can occur in experiments with combinations of many biological samples and perturbations. Keeping our hypergeometric model, we take inspiration from the Good-Turing frequency estimation,

which can be used to estimate the frequency of previously unseen species<sup>44</sup>. Good-Turing estimation states that given a transcripts belonging to gene  $i$  is found  $r$  times in a pool of transcripts containing a total of  $N$  transcripts and the number of genes that are found  $r$  times is  $n_r$ , we should estimate the frequency of gene A as:

$$\frac{1}{N}(r+1)\frac{n_{r+1}}{n_r} \quad (2.11)$$

We apply this equation to single-cell data at the biological sample level, bringing us to the final mean estimator:

$$\begin{aligned} r_g &= \sum_c Y_{cg}, \text{ count of gene } g \text{ in the sample} \\ n_r &= \sum_g \mathbb{1}(r_g = r), \text{ number of genes with count } r \\ r_g^* &= (r_g + 1)\frac{n_{r_g+1}}{n_{r_g}} \\ \hat{\mu}_{g,memento} &= \hat{\mathbb{E}}[X_{cg}] = \frac{1}{\sum_c N_c} r_g^* \end{aligned} \quad (2.12)$$

## Estimating cell sizes by trimming variable genes

The  $N_c q_c$  values that appear in the HG estimator equations above refer to the cell size, which serves as a normalization factor for each cell. These constants serve to ensure that even if the proportions of transcripts captured vary across cells, the estimates would not be affected by this technical source of noise. We can decompose  $N_c q_c$  into two components: a constant  $n_{umi}$  and  $\gamma_c$  so that  $N_c q_c = n_{umi} \gamma_c$ . The simplest way of estimating  $\gamma_c$  is to first compute  $n_{umi} = \frac{1}{n_{cells}} \sum_c \mathbf{1}^T \mathbf{Y}_c$ , and setting  $\gamma_c = \frac{1}{n_{umi}} \mathbf{1}^T \mathbf{Y}_c$ , performing a total count normalization.

This is how the Poisson estimators presented in Zhang et al.'s work estimated the cell sizes. In *memento*, we provide an alternate method by first computing residual variances across all

cells in a dataset with total count normalization, and trimming off genes that exhibit high variability. This approach assumes that most genes in the dataset should not be differentially expressed, and the least variable genes are appropriate to be used in normalization. This idea of using non-DE genes have been used in other methods, such as<sup>45-47</sup>. By default, **memento** uses 10% of the least variable genes. After gene set  $G^*$  is formed by trimming variable genes, we compute  $\gamma_c$  with:

$$\gamma_c = \frac{\delta + \sum_{g \in G^*} Y_{cg}}{\delta + \frac{1}{n_{cells}} \sum_c \sum_{g \in G^*} Y_{cg}}$$

The  $\delta$  value here serves as a regularization factor in estimating cell sizes; when this value is high, it would indicate the dataset does not need a size factor normalization (sampling is truly constant across cells, such as when sequencing to saturation). By default, **memento** uses  $\text{median}(\sum_{g \in G^*} Y_{cg})$  over cells  $c$  as the  $\delta$  value.

It is important to note that there are more sophisticated normalization methods that exist in literature<sup>48</sup>. **memento** can readily incorporate these alternative methods of computing cell sizes into its pipeline.

## Computing the residual variance

Mean and variance in scRNA-seq data is generally highly correlated and measuring variability of expression must account for this correlation. BASiCs accounts for this dependence by performing nonlinear regression with many components between the fitted mean and overdispersion parameters<sup>27</sup>. Instead of fitting a negative binomial distribution then regressing out the mean from the overdispersion parameter, we simply take the estimated true mean and variances and fit a simple polynomial regression. We use a single fitted polynomial (default degree 2) for all genes of a given group of cells, defined by cell type, experimental condition, or batch. We find that even this simple regression is able to

largely remove the mean-variance dependence present in scRNA-seq data.

## Efficient bootstrapping by exploiting data sparsity

Typically, generating confidence intervals and computing p-values for hypothesis testing make certain assumptions on both the distribution of the data as well as the estimator itself. For example, to compute p-values for the coefficients of a linear regression model, we typically assume that the data is normally distributed and the sampling distribution of the coefficients are also normal. In the setting of scRNA-seq, our estimators allow for measurement of the average, variability, and gene correlation without making any assumptions about the distribution of expressed transcript counts. However, it is difficult to compute analytical confidence intervals for our estimators without assuming anything about the data itself and the sampling distributions of our estimates.

Bootstrapping is a procedure for estimating the sampling distribution of any arbitrary statistic without making large assumptions on the data generating process<sup>41</sup>. In **memento**, we propose a strategy to perform bootstrapping in scRNA-seq data in an extremely efficient manner. Specifically, in a dataset for a single gene with  $N$  cells  $x_1, x_2, x_3, \dots, x_N$ , we can model the number of appearance of each observation as a multinomial distribution with  $\text{Multinomial}(N, \frac{1}{N} \dots \frac{1}{N})$ . If there are  $K$  unique counts with  $n_k$  cells each, we can re-write the resampling distribution as  $\text{Multinomial}(N, \frac{n_1}{N} \dots \frac{n_K}{N})$ .

When considering normalized transcript abundances, we must account for the total number of transcripts in each cell ( $N_c$ ). While this would technically create a different  $N_c$  for each cell and make our scheme less useful, a strategy binning  $N_c$ s across cells into a small number of discrete bins well-approximates the true bootstrap distribution of parameters. Through simulations, we show that as the number of bins increase, we show that the true bootstrap distribution and the approximate bootstrap distributions are nearly identical (**Fig. S2.7**).

## Hypothesis testing and extension to account for replicates in multiplexed scRNA-seq experiments

Consider a scenario with two groups of cells A and B, and we computed the parameter of interest  $t$  for each group and computed  $\Delta t$  as their difference.  $t$  would depend on the type of test we would like to perform; we would compute the mean, residual variance, and correlation to test for differences in the averages, variability, and coexpression respectively. We then perform bootstrapping with  $B$  iterations to generate a sampling distribution for the test statistic  $\Delta t$ , from  $\Delta t_1$  to  $\Delta t_B$ . If we wished to test for the alternative hypothesis of H1:  $\Delta t \neq 0$  against the null H0:  $\Delta t = 0$ , we first generate the null distribution by subtracting  $\Delta t$  from  $\Delta t_1, \dots, \Delta t_B$  to form  $\Delta t_1^*, \dots, \Delta t_B^*$ , similar to the strategy laid out in<sup>41</sup>. We can then compute the achieved significance level (ASL) for that test as

$$\text{ASL} = \begin{cases} \frac{2}{B} \sum_{i=1}^B \mathbb{1}(\Delta t > \Delta t_i^*) & \text{if } \Delta t \geq 0 \\ \frac{2}{B} \sum_{i=1}^B \mathbb{1}(\Delta t < \Delta t_i^*) & \text{if } \Delta t < 0 \end{cases}$$

There has been an increasing trend to generate scRNA-seq data with replicates (e.g. different individuals), especially with multiplexed workflows. Consider an experiment with two conditions and  $n$  replicates. Then, we propose a meta-analysis framework where we first group the cells into  $2n$  groups and perform a meta-regression with  $2n$  observations:

$$\begin{bmatrix} \ln \mu_1 \\ \ln \mu_2 \\ \vdots \\ \ln \mu_{2n-1} \\ \ln \mu_{2n} \end{bmatrix}, \begin{bmatrix} \ln \tilde{\sigma}_1 \\ \ln \tilde{\sigma}_2 \\ \vdots \\ \ln \tilde{\sigma}_{2n-1} \\ \ln \tilde{\sigma}_{2n} \end{bmatrix}, \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{2n-1} \\ \rho_{2n} \end{bmatrix} \sim \beta \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_{2n-1} \\ W_{2n} \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad (2.13)$$



where  $\mu_i$ ,  $\tilde{\sigma}_i$ , and  $\rho_i$  refer to the estimated mean, residual variance, and correlation computed in the  $i^{\text{th}}$  replicate and  $W_i$  refers to the condition. Then, we can bootstrap the regression coefficients  $B$  times to yield the original statistic  $\hat{\beta}$  and bootstrap statistics  $\hat{\beta}_1, \dots, \hat{\beta}_B$ . Then, similar to the non-replicated case, we can generate the null distribution  $\hat{\beta}_1^*, \dots, \hat{\beta}_B^*$  by subtracting  $\hat{\beta}$  from  $\hat{\beta}_1, \dots, \hat{\beta}_B$ . We can further compute the ASL with:

$$\text{ASL} = \begin{cases} \frac{2}{B} \sum_{i=1}^B \mathbb{1}(\hat{\beta} > \hat{\beta}_i^*) & \text{if } \hat{\beta} \geq 0 \\ \frac{2}{B} \sum_{i=1}^B \mathbb{1}(\hat{\beta} < \hat{\beta}_i^*) & \text{if } \hat{\beta} < 0 \end{cases}$$

Alternatively, the null distribution  $\hat{\beta}_i^*$  can be approximated as a normal distribution  $N(0, \sigma^{*2})$ , and the significance level can be calculated as  $2(1 - \Phi(|\hat{\beta}|/\sigma^*))$ .

This framework can easily be extended to incorporating many covariates, including batch variables and interactions between variables of interest by introducing additional covariates into the model in equation 2.13, by providing additional columns aside from the treatment variables  $W$ . Information at the level of groups of cells, such as age, sex, genotypes can be incorporated similar to how they would be incorporated into a generalized linear model. Incorporating covariates at the single cell level is currently not handled by `memento`.

As a technical aside, we note that this procedure for computing the ASL assumes that the sampling distribution of the test statistic of interest is translation invariant<sup>41</sup>. Through extensive simulations, we confirm that for the test statistics we consider in `memento`, this procedure yields well-calibrated results under the null hypothesis (**Fig. S2.10**). If custom test statistics are used, it is important to check for the calibration of hypothesis test results. `memento` also has the option to compute p-values assuming that the sampling distribution of the effect size is normal with unknown variance that is estimated using the bootstrap, useful for speeding up hypothesis tests. For this work, this approximation was only used for analyzing the effect of natural variation (**Fig. 3.3**).

## Pre-processing the rIFNB1 PBMC dataset

We used the original clustering and tSNE visualization of the rIFNB1 dataset<sup>15</sup> from the data deposited in the Gene Expression Omnibus under the accession number GSE96583. Further details on the pre-processing of this dataset can be found in the original paper<sup>15</sup>.

For all analysis, we selected genes where the mean observed expression  $\mathbb{E}[Y_{cg}] = 0.07$ , which was the reliability limit for this experiment. More details on the reliability limit can be found in<sup>37</sup>. This value was computed from the reported UMI capture efficiency of 10X Chromium V1 and well as the sequencing saturation of this experiment, which was around 90%<sup>15</sup>.

## Extracting mean and variance from scRNA-seq data for simulation

We used the PBMC dataset<sup>15</sup> to serve as a basis for the simulation. We wanted to simulate single cell RNA profiles that have a distribution of means and variances that are within the realistic regime of scRNA-seq. To accomplish this, we estimated the mean, variance, and correlation of 5000 highest expressed genes using the memento estimators from the CD4+ T cells. These values were then set as the ground truth parameters for the simulation. We used two sets of ground truth parameters - one estimated from cells without stimulation ( $m_{unstim}, v_{unstim}$ ), and one from cells stimulated with IFN-B ( $m_{stim}, v_{stim}$ ).

## Simulating transcriptomes with given means, variances, and gene-gene correlations

Given a vector of desired means ( $m$ ) and variances ( $v$ ), we first calculated the dispersion by using moment calculations dispersion =  $\frac{v-m}{m^2}$ . To generate transcriptomes with ground truth correlations, we took the following steps:

1. Generating correlated zero mean, unit variance Gaussian samples using the ground truth correlation parameter

2. Computing the copula by taking the inverse Gaussian CDF of each point
3. Generating the marginal distribution by taking evaluating the negative binomial point percent point function with the specified mean and dispersion vectors previously calculated.

This process implements a Gaussian copula method for generating multivariate samples from a joint distribution with a specified correlation matrix and negative binomial marginal distributions. Note that *memento* does not make any assumptions about the underlying distribution, and the negative binomial was used here to be consistent with past strategies for simulating scRNA-seq data<sup>27</sup>.

After the "true" transcriptomes are simulated, we sample the transcripts with the hypergeometric distribution with a overall capture efficiency  $q$  (combining sampling from library preparation and sequencing).

## Comparing *memento*, BASiCS, and scHOT for estimation

To generate Figure 2.22A, we generated transcriptomes using  $m_{unstim}$ ,  $v_{unstim}$  estimated above from a real scRNA-seq dataset, and a correlation matrix  $C$  sampled from `make_psd` function from the `scikit-learn` package, while varying the overall capture efficiency  $q_{real}$ . We estimated the means and correlations from *memento* (hypergeometric), *memento* ( $q=0$ ), naïve estimators. We estimated the variances using *memento* (hypergeometric), *memento* ( $q=0$ ), naïve, and BASiCS estimators. We calculated the variance using the dispersion estimates from BASiCS output by using the mean-variance relationship for the negative binomial distribution. Because we cannot directly compute the residual variance in the smFISH data, we used the coefficient of variation in place of residual variance for this analysis. This process was repeated 20 times for each value of overall capture efficiency to generate confidence intervals. We used simulations of 10 cells for the mean and 100 cells for variability

and gene correlation.

## Simulating genes with differential mean, variability, and coexpression

To compare the performance of `memento` for differential expression against other methods in a realistic, complex experimental settings, we used a hierarchical simulation with hierarchical generation.

For differential mean, we first computed  $\Delta m = \log(m_{stim}) - \log(m_{unstim})$ . To designate ground truth DM genes, we set any elements of  $\Delta m$  lower than 0.1 to 0. We simulated data with 2 replicates, creating 4 total groups of cells: unstimulated replicate 1, stimulated replicate 1, unstimulated replicate 2, unstimulated replicate 2. We generated the four sets of mean vectors as:

$$m_{1,unstim} = N(m_{unstim}, 0.25)$$

$$m_{2,unstim} = N(m_{unstim}, 0.25)$$

$$m_{1,stim} = m_{1,unstim} + \Delta m + N(0, 0.25)$$

$$m_{2,stim} = m_{2,unstim} + \Delta m + N(0.25)$$

These mean vectors represent baseline variations that exist across replicates (such as individuals) and heterogenous treatment effects (cells from different replicates may not respond in an identical way). We then simulated varying numbers of cells (1000, 1000, 1100, 1100) to emulate varying sample sizes from each replicate using the procedure described in Simulating transcriptomes with given means, variances, and gene-gene correlations. For differential mean simulations, we set all variances as  $v_{unstim}$  and induced no correlations between genes.

For differential variability, we first computed  $\Delta v = \log(v_{stim}) - \log(v_{unstim})$ . To designate ground truth DV genes, we set any elements of  $\Delta v$  lower than 0.1 to 0. We simulated data with 2 replicates, creating 4 total groups of cells: unstimulated replicate 1, stimulated

replicate 1, unstimulated replicate 2, unstimulated replicate 2. We generated the four sets of variance vectors as:

$$v_{1,unstim} = N(v_{unstim}, 0.25)$$

$$v_{2,unstim} = N(v_{unstim}, 0.25)$$

$$v_{1,stim} = v_{1,unstim} + \Delta v + N(0, 0.25)$$

$$v_{2,stim} = v_{2,unstim} + \Delta v + N(0.25)$$

These variance vectors represent baseline variations that exist across replicates (such as individuals) and heterogeneous treatment effects (cells from different replicates may not respond in an identical way). We then simulated varying numbers of cells (500, 500, 700, 700) to emulate varying sample sizes from each replicate using the procedure described in Simulating transcriptomes with given means, variances, and gene-gene correlations. For differential variability simulations, we set the mean vectors in the same way as simulated differential mean.

For differential correlation, we followed an similar approach as differential mean and variability, but we generated  $\Delta corr$  by subtracting two random correlation matrices generated with `make_psd` function in the scikit-learn model.

Similar to the simulations performed to compare estimation performance, we sample the "true" transcriptome's transcripts with the hypergeometric distribution with a overall capture efficiency  $q$  (combining sampling from library preparation and sequencing).

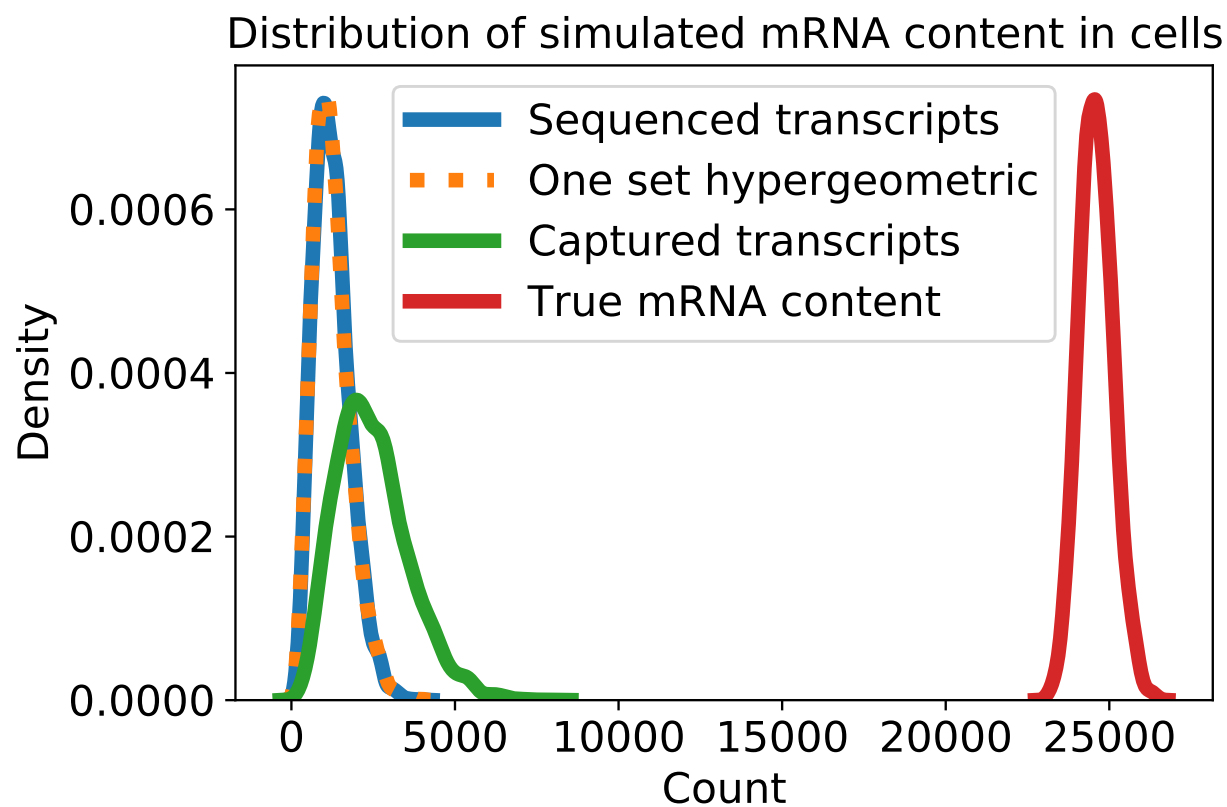
## Comparing DE methods, BASiCS, and scHOT for differential mean, variability, and correlation

For comparing memento to established differential mean expression methods, we used the same parameters used by Squair et al (2022). For BASiCS, we ran the method with the no-spike in mode and the regression modes. For scHOT, we used the default parameters

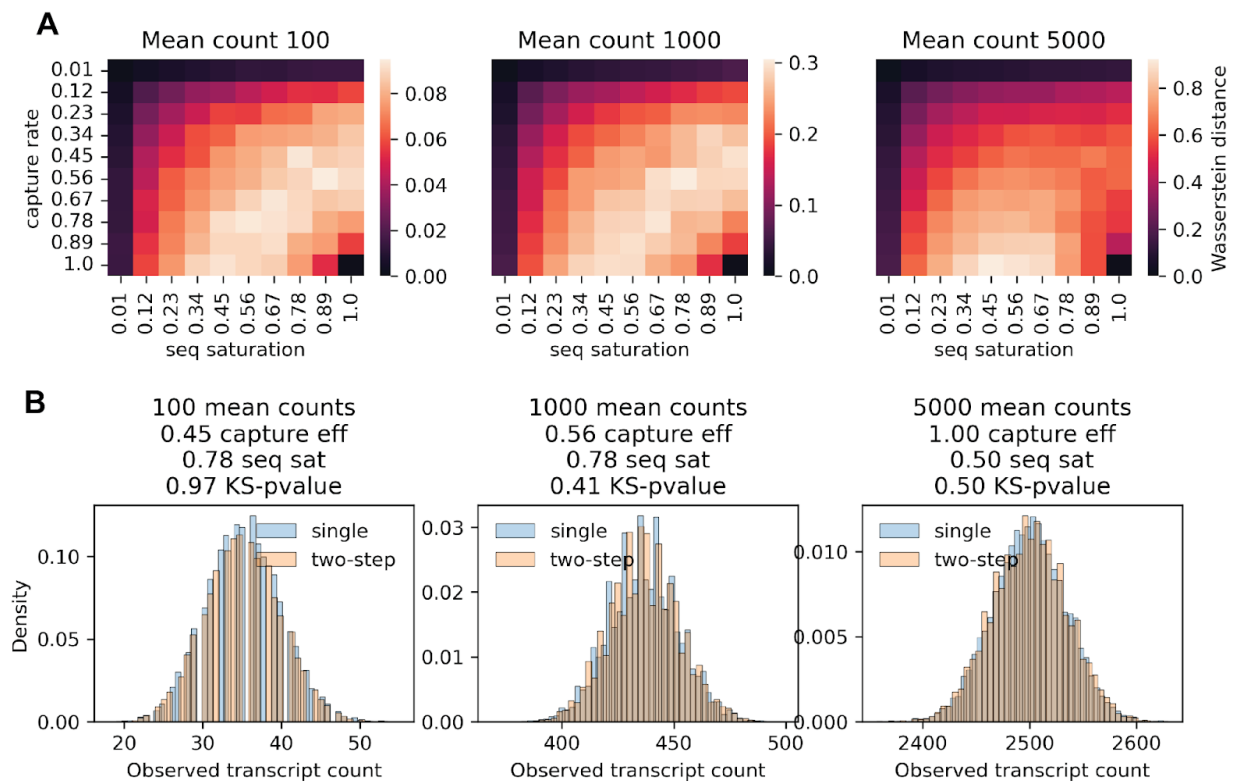
with binary mask for the “pseudotime” to compute parameters across.

We performed 20 repeated simulations at each cell count across varying cell counts to generate Figure 2.22C.

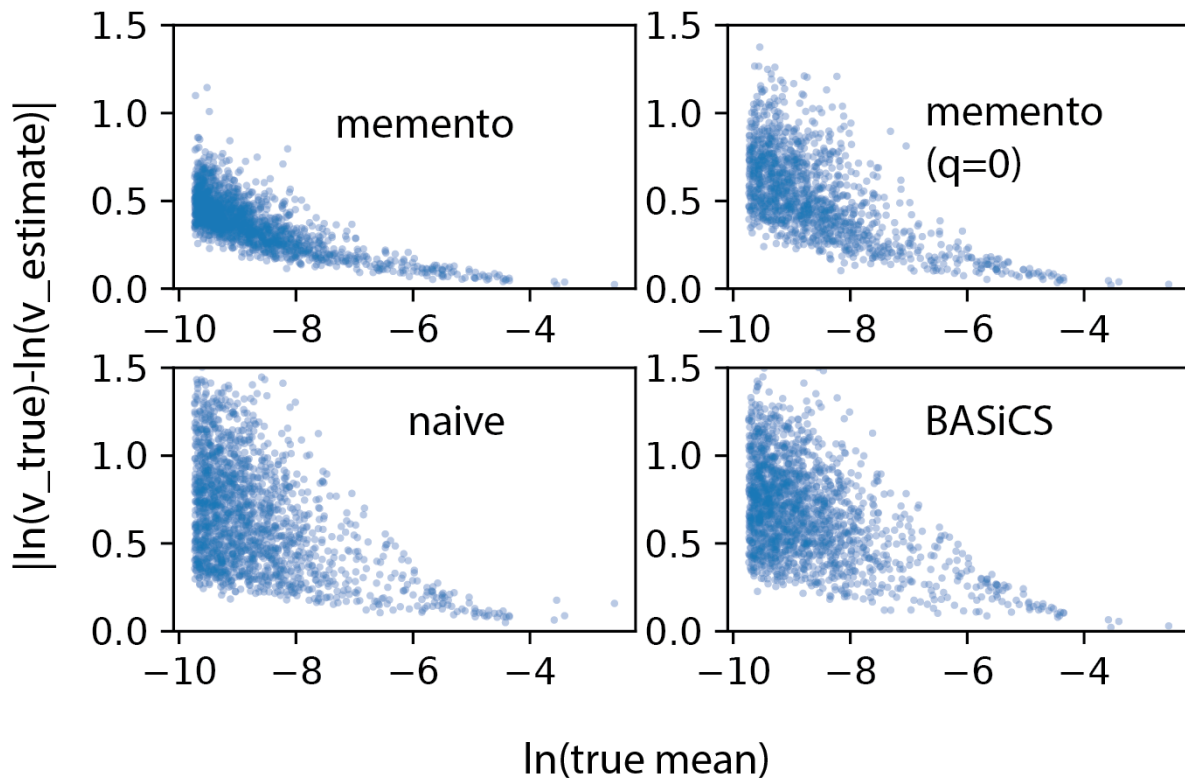
## 2.6 Supplementary figures



**Figure S2.1:** Single step of hypergeometric sampling well approximates the compound sampling process from capture and sequencing.

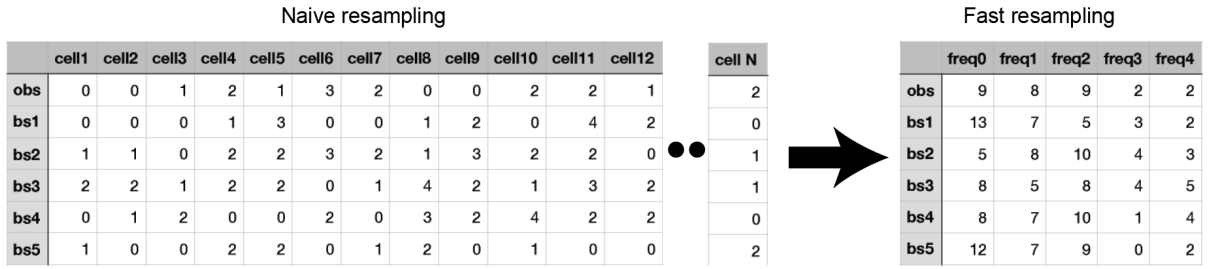


**Figure S2.2:** Characterizing the effect of approximating the two-step sampling in a single hypergeometric sampling step. **(A)** Heatmap of Wasserstein distance between distributions resulting from various capture efficiency and sequencing saturation. **(B)** Histogram of the two different sampling processes at a single capture efficiency and sequencing saturation (one of the tiles in panel A).

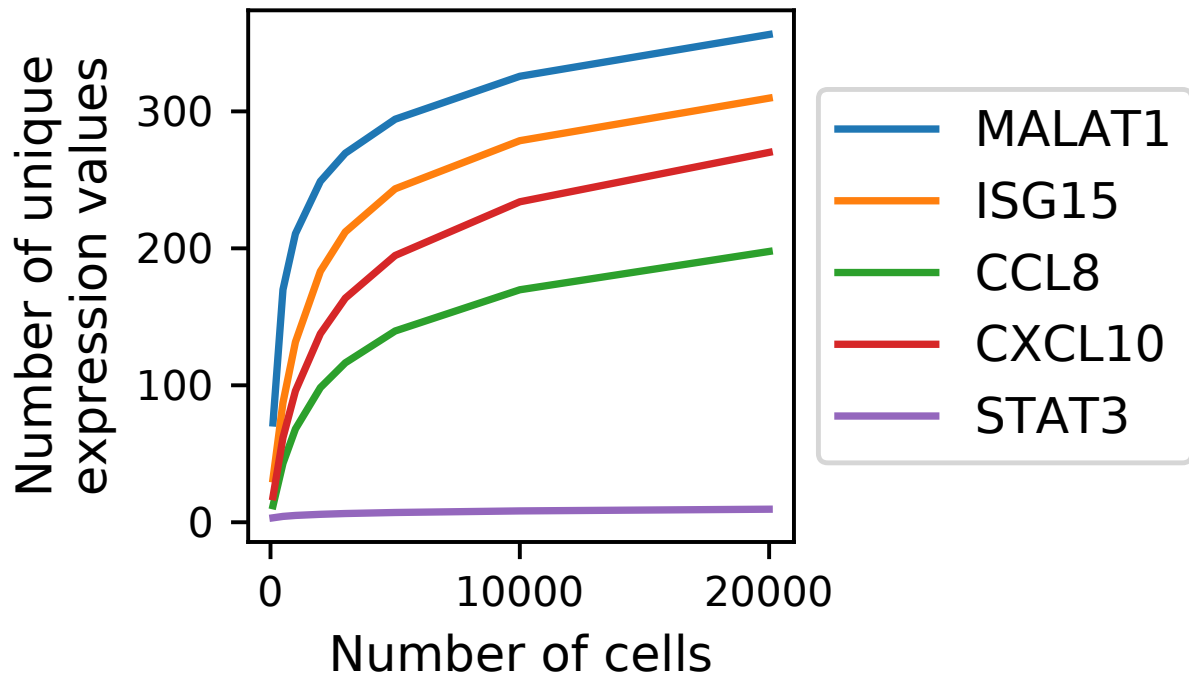


**Figure S2.3:** Each point represents a gene in simulation. Y-axis is error of variance estimation ( $|\ln(v_{true}) - \ln(v_{estimated})|$ ) and x-axis is the gene's true simulated mean.

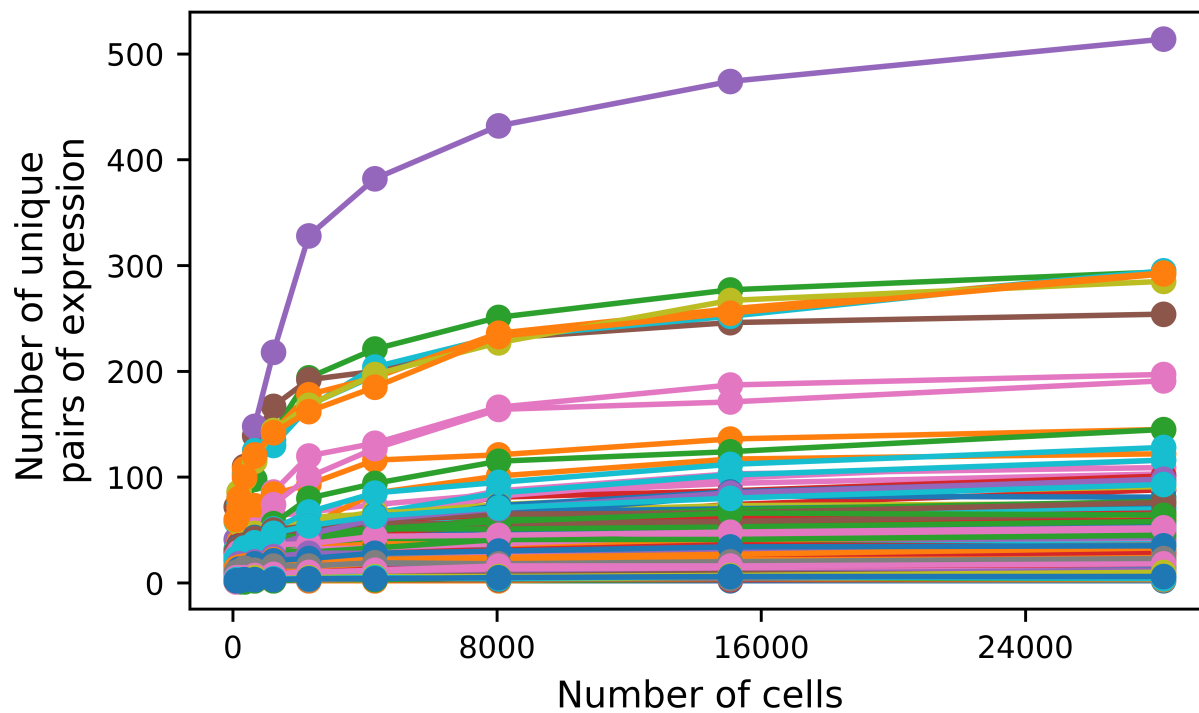




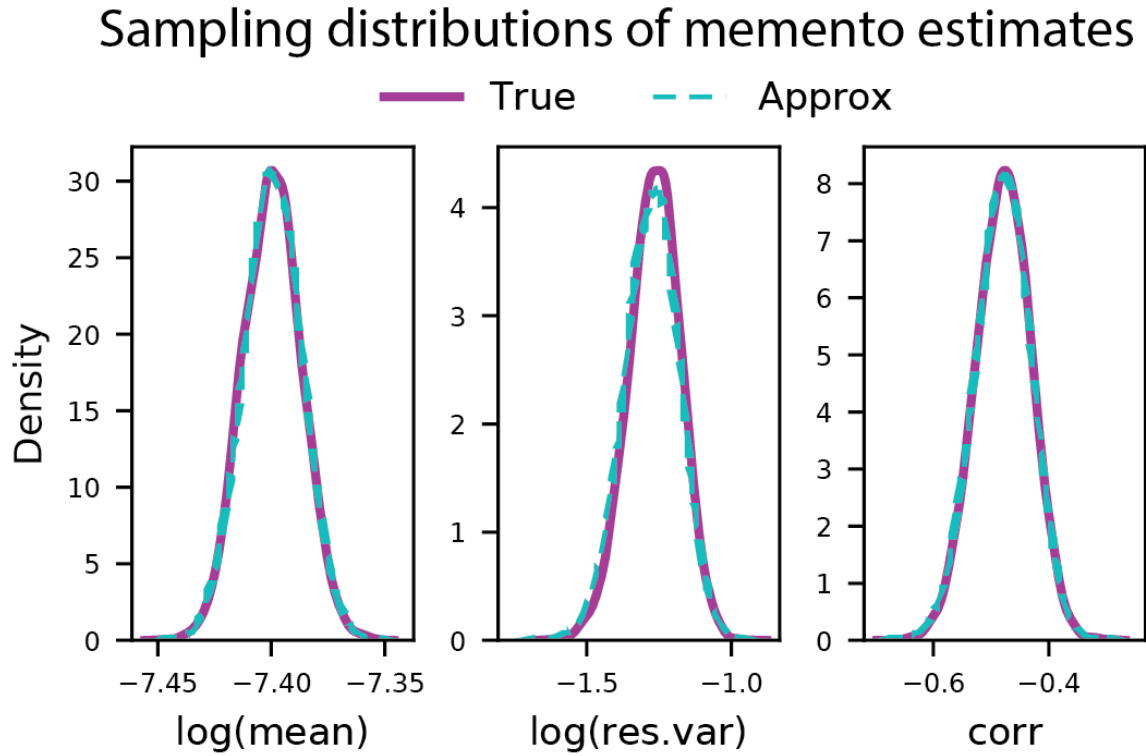
**Figure S2.4:** Conceptual diagram for resampling frequencies rather than expression counts.



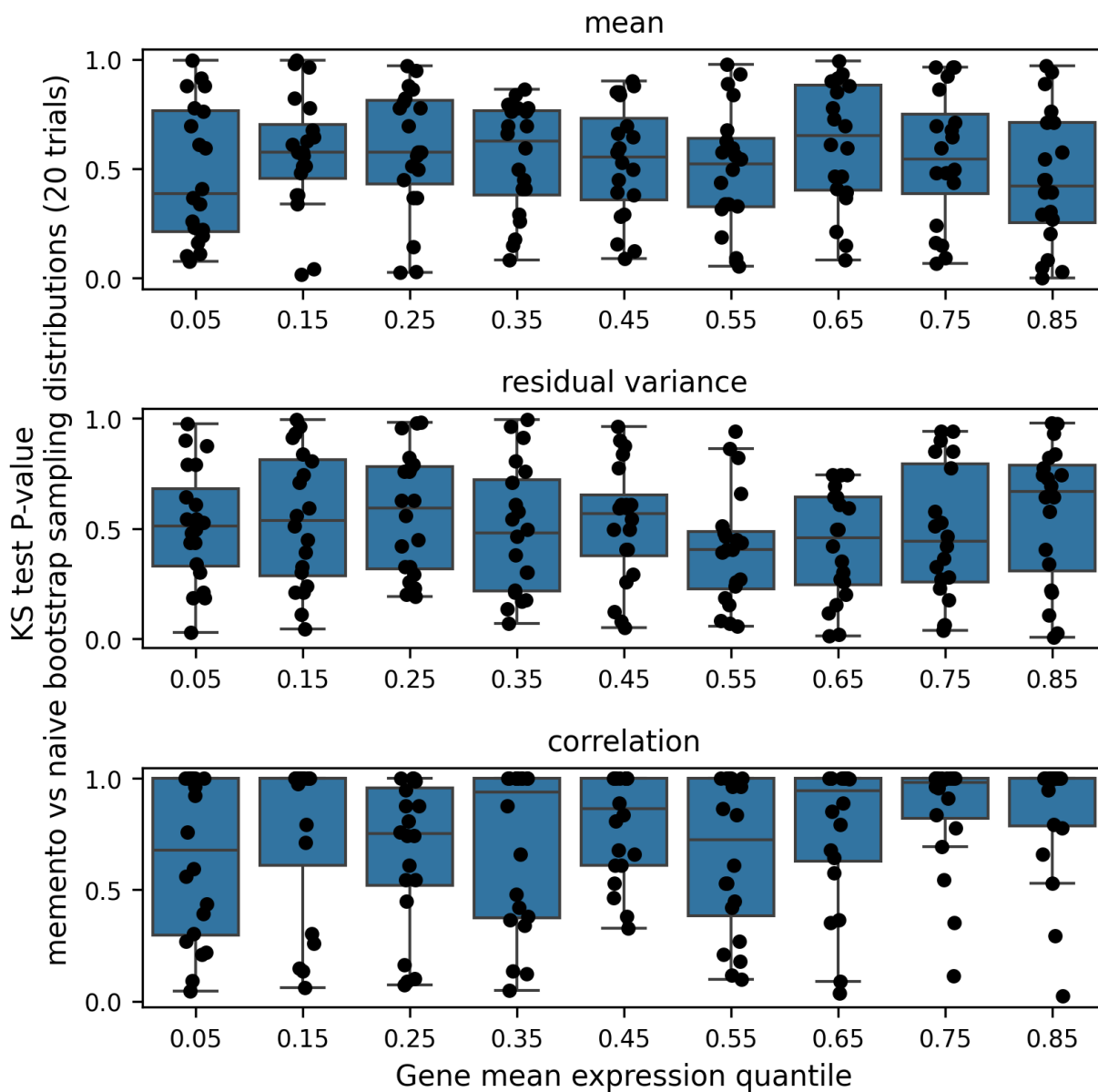
**Figure S2.5:** Number of unique transcript counts (y-axis) against the number of cells present in the dataset (x-axis)



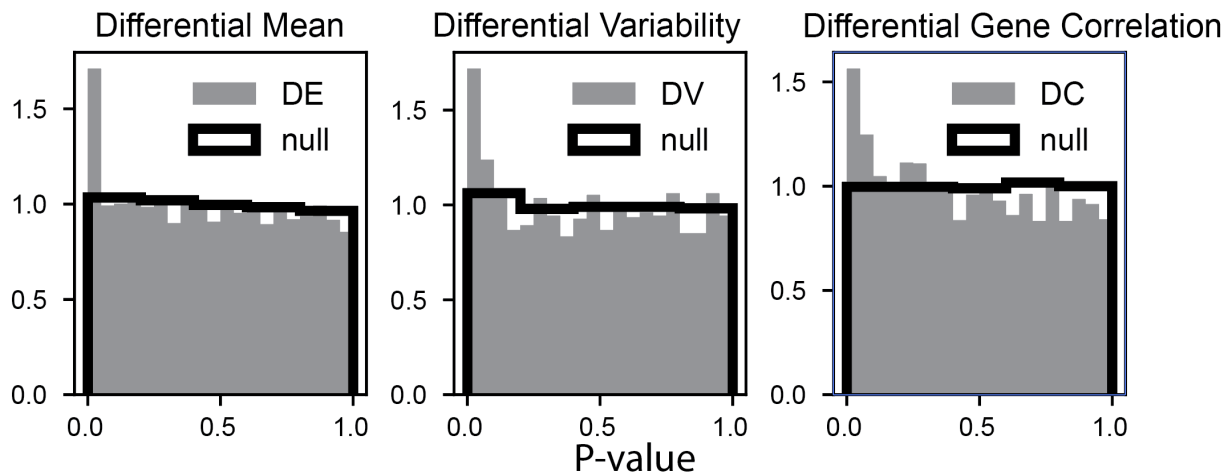
**Figure S2.6:** Number of unique pairs of genes for randomly selected pairs in the IFN-B dataset.



**Figure S2.7:** Efficient bootstrap in memento vs full bootstrap. The sample distributions of  $\log(\text{mean})$ ,  $\log(\text{residual variance})$  and correlation for a representative gene and a pair of genes, respectively.



**Figure S2.8:** Comprehensive comparison of *memento*'s bootstrap with the naive bootstrap. Y-axis is KS-test P-value from comparing the *memento* versus naive bootstrap distributions for the mean (top), residual variance (middle), and correlation (bottom). X-axis is the gene's mean expression quantile.



**Figure S2.9:** Representative example of P-values computed from memento

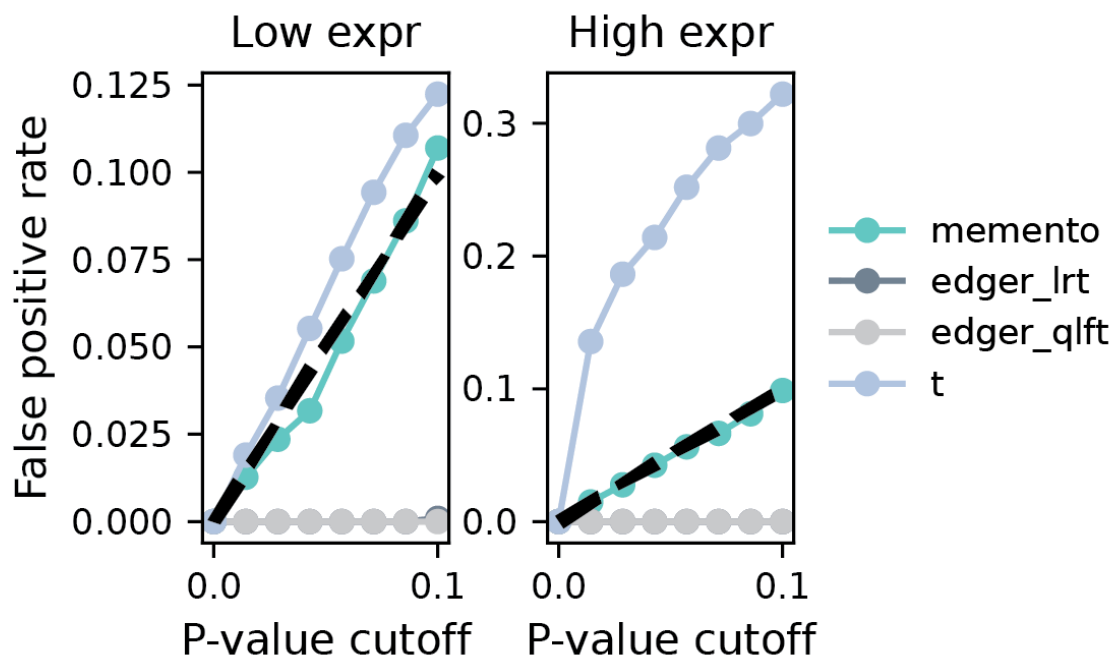
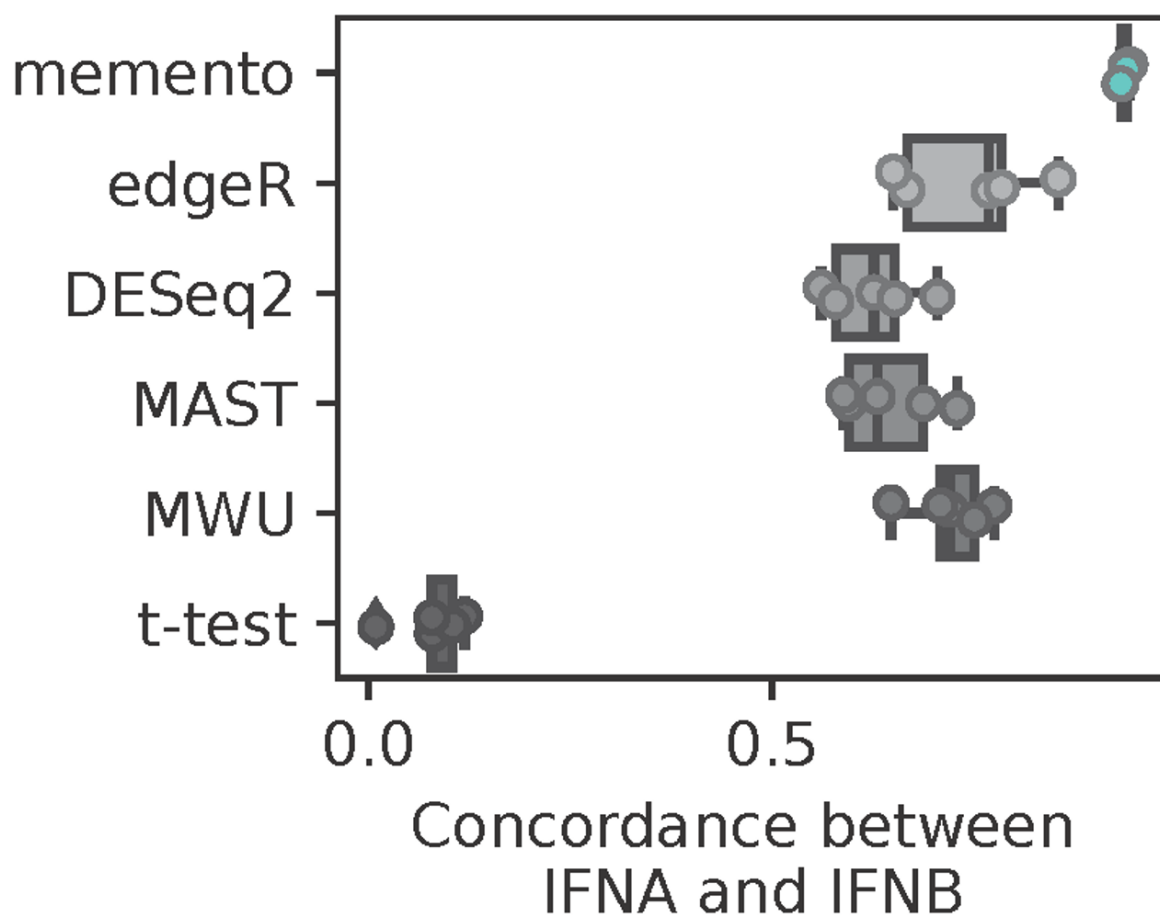


Figure S2.10: False discovery rates in simulation for lowly and highly expressed genes



**Figure S2.11:** Concordance of DE genes between IFN-A and IFN-B across various methods



# Chapter 3

## Generalized differential expression across natural variation and experimental conditions

### 3.1 Differential variability and gene correlation in response to exogenous interferon

Interferons, while being potent cytokines that promote antiviral immunity, also play a part in the pathogenesis of inflammatory and autoimmune diseases<sup>49</sup>. Their action - inducing gene expression via autocrine and paracrine signaling - is well-documented, yet the heterogeneity of transcriptomic responses in stimulated cells remain unexplored. Using **memento**, we investigated the impact of interferon stimulation on the distribution of gene expression in human tracheal epithelial cells (HTECs). We used multiplexed single-cell RNA-sequencing to analyze 69,958 HTECs from two healthy donors, exploring conditions including unstimulated control and stimulation with various interferons: type-1 (IFN- $\alpha$ , IFN- $\beta$ ), type-2 (IFN- $\gamma$ ), and type-3 (IFN- $\lambda$ ). Analyses were conducted at several post-stimulation timepoints: 3, 6, 9, 24, and 48 hours. Dimensionality reduction, nearest neighbor identification, and Leiden clustering yielded 7 identifiable cell types, visualized using uniform manifold approximation

and projection (UMAP): neuroendocrine cells, ionocytes, tuft cells, basal cells, basal/club cells, goblet cells, and ciliated cells (**Fig. 3.1A**). Our subsequent analysis focused solely on ciliated cells, which are known to be the primary target of viral infections including SARS-CoV2 and are recognized for their robust interferon response<sup>50-52</sup>.

We identified 5,018 genes exhibiting differential mean expression (DMGs, FDR < 0.01) between unstimulated ciliated cells and those stimulated by any of four interferons at 6 hours. A comparative analysis revealed that IFN- $\alpha$  shares similar mean expression changes with IFN- $\beta$  and IFN- $\lambda$  in differential mean expression ( $\rho = 0.96$ ). In contrast, comparing to IFN- $\gamma$  highlighted distinctions in differential mean expression for both type-1 and type-2 interferon-specific genes ( $\rho = 0.70$ , **Fig. 3.1B**). Herein, we define genes that are upregulated in response to any interferon as interferon-stimulated genes (ISGs). Hierarchical clustering of the ISGs at the 6-hour timepoint revealed a dynamic transcriptomic response, shared across interferons, that included the early induction of MHC class II genes and a distinct gene cluster, comprising *PLAAT2*, *BTN3A1*, and *DUOX2* (**Fig. 3.1C**). Furthermore, we identified patterns specific to each interferon, exemplified by a subset of canonical ISGs (*IFI2*, *IFITM2*, and *ISG15*) that not only exhibited continual increase in response to IFN- $\lambda$  but also sustained elevated levels throughout the experimental timeline for type-1 interferons (**Fig. 3.1C**). Interestingly, some genes that were more strongly expressed in one of the interferons (e.g., the MHC class II genes and IFN- $\gamma$ ) showed similar temporal behavior across the other interferons, suggesting both unique and shared regulatory mechanisms.

While the analysis of differential mean expression revealed the induction of canonical and non-canonical ISGs, it did not decipher whether these genes were subject to the same transcriptional regulatory control. To delineate the interferon gene correlation network and its subcomponents, we used *memento* to compare correlations between ISG pairs across various stimulations and timepoints (**Fig. 3.1D**). Agglomerative clustering of the resulting gene correlation matrix revealed distinct ISG subsets in response to IFN- $\beta$ , forming clusters

in unstimulated cells, stimulated cells, or both—distinctions that were not discernible through differential mean analysis alone. For example, canonical ISGs including *MX1*, *OAS1*, and *IFI6*, maintained high correlation even without exogenous interferon presence (**Fig. 3.1D**, cyan nodes). Upon IFN- $\beta$  stimulation, the correlation network, initially consisting of canonical ISGs, expanded to include non-canonical ISGs, such as the MHC Class I molecules and other genes associated with antigen presentation, which were not correlated in unstimulated cells (**Fig. 3.1D**, magenta nodes). Remarkably, more differentially correlated gene pairs (DCGs, FDR < 0.1) were found among non-canonical ISGs (860 DCGs, 34% of total pairs) than canonical ISGs (421 DCGs, 16% of total pairs). In addition, we found that the increase in correlation was not explained by the increase in the mean expression of those genes (**Fig. S3.1**).

We hypothesized that canonical ISGs display correlation in unstimulated cells due to the sensing of tonic interferon and the coordinated induction of ISGs within a select group of cells. Tonic interferon signaling has been described to induce a natural gradient of ISG expression across cells<sup>53,54</sup>, and plays an important role in viral defense<sup>54</sup>, immune cell homeostasis, and autoimmunity<sup>53</sup>. Within our dataset, canonical ISGs exhibited greater variability compared to non-canonical ISGs in unstimulated cells (**Fig. 3.1E**), aligning with previously documented differences in expression variability between cytokines and non-cytokines (**Fig. S3.2**)<sup>55</sup>. Out of the 761 differentially variable genes (DVGs, FDR < 0.1) identified using *memento* between unstimulated ciliated cells and those stimulated by any of the four interferons at 6 hours, 394 were highly variable in unstimulated cells (FDR < 0.005) and were enriched for ISGs (GSEA Interferon alpha/beta signaling Adjusted  $P = 3.35 \times 10^{-12}$ ), including *IFIT1*, *IFIT3*, and *MX1*.

We next compared the tonic sensitivity of each canonical and non-canonical ISGs, estimated as the fold-change (FC) in gene expression between macrophages from *IFNAR* knockout and wild-type mice without exogenous interferon<sup>56</sup>. This analysis revealed that canonical ISGs

are significantly more sensitive to tonic interferon than non-canonical ISGs ( $P < 2.73 \times 10^{-10}$ ), **Fig. 3.1F**). Notably, upon stimulation with IFN- $\beta$  (and, to a lesser extent with IFN- $\gamma$ ), the variability of substantial proportion of canonical ISGs diminished (78% and 39%, respectively) (**Fig. 3.1G**, FDR  $< 0.1$ ), implying that exogenous stimulation might homogenize the cellular environment, removing the effects of heterogeneous response to tonic interferon.

Our findings demonstrate the applicability of *memento* for the comparison of gene expression distributions to reveal novel transcriptional regulatory modalities influenced by extracellular interferon. Within HTECs, our discoveries encompass: 1) a core network of canonical ISGs, exhibiting highly variable and correlated gene expression in unstimulated cells, attributable to tonic interferon signaling; 2) the homogenization of canonical ISGs upon exogenous interferon exposure, leading to diminished variability; and 3) a network of non-canonical ISGs, which are exclusively modulated in response to exogenous and not tonic interferon.

## 3.2 Differential expression analysis of perturbed CD4<sup>+</sup> T cells maps gene regulatory networks in T cell activation

Integrating CRISPR-Cas9-mediated genomic perturbations with scRNA-seq profiling creates new opportunities for conducting forward genetics screens in diverse in vitro systems.

Utilizing *memento*, we analyzed 173,000 CRISPR-Cas9 perturbed CD4<sup>+</sup> T cells to map transcriptional regulatory networks modulating the activation and polarization of human CD4<sup>+</sup> T cells. Cells were perturbed using pooled sgRNA lentiviral infection with Cas9 protein electroporation (SLICE)<sup>57</sup>, followed by multiplexed single-cell RNA-sequencing (mux-seq). Utilizing a set of 280 sgRNAs, we targeted 140 transcriptional regulators (TRs), chosen for their high expression (within the top quartile from bulk RNA-seq) or the

differential accessibility of their binding sites (as detected by bulk ATAC-seq) in activated CD4<sup>+</sup> T cells<sup>58</sup> (**Fig. 3.2A**). After Cas9 electroporation, and multiple rounds of selection and proliferation, activated CD4<sup>+</sup> T cells from 9 donors were profiled using mux-seq.

To evaluate the cutting efficiency of each sgRNA, we conducted targeted amplification sequencing of 268 out of 280 loci in both the sgRNA pool and the DNA of edited cells from each donor. The mean cutting efficiency across 268 sgRNAs, defined as the fraction the coverage of edited cells at the target locus to the coverage of its respective sgRNA in the pool, was established at 21%, with a standard deviation of 15% (**Fig. S3.3**). Fourteen sgRNAs, exhibiting cutting efficiencies below 2.0% (standard deviation 1.7%; z-score,  $P < 0.05$ ), were designated as uncut negative controls (WT). The robustness and efficacy of our screen were substantiated through two quality control analyses. First, we utilized *memento* to confirm a significant downregulation of target genes in cells transduced with the respective sgRNA (**Fig. 3.2B**). Second, a higher correlation in average gene expression was observed between either WT cells ( $\rho = 0.50$ ) or cells transduced with sgRNAs targeting identical genes ( $\rho = 0.44$ ), as compared to cells transduced with sgRNAs targeting two distinct genes ( $\rho = 0$ ; KS-test  $P < 2.2 \times 10^{-16}$  for both; **Fig. S3.4**).

Utilizing *memento*, we identified 7,641 genes (FDR  $< 0.05$ ) with differential mean expression (DMGs) when contrasting WT cells against cells perturbed by at least one sgRNA.

Hierarchical clustering revealed groups of sgRNAs exerting similar transcriptomic effects and gene groupings similarly responsive to such perturbations (**Fig. 3.2C**). We identified five clusters of DMGs distinctly associated with ribosomes (FDR  $< 5.35 \times 10^{-24}$ ), cytotoxicity (FDR  $< 0.014$ ), antigen presentation (FDR  $< 0.0011$ ), and proliferation (FDR  $< 0.001$ ).

Moreover, the pairwise correlation matrix of DMGs, as computed using *memento*, revealed additional sub-clusters within each of the initial five DMG clusters, persisting in both WT and perturbed cells (**Fig. 3.2C**). Intriguingly, while antigen processing genes' mean expression is modulated by a shared set of transcriptional regulators, a subset of MHC class

II genes—namely *HLA-DPA1*, *HLA-DRA*, *HLA-DRB1*, and *HLA-DPB1*—exhibited strong correlation, suggesting that their expression may be controlled by additional *trans* regulators.

In exploring the utility of *memento* for detecting alterations in gene correlations, we hypothesized that identification of genetic interactions between transcriptional regulators might be achievable without necessitating combinatorial perturbations. To test this hypothesis, we performed a genetic interaction analysis focused on pairs consisting of DMGs and their transcriptional regulators, referred to as transcriptional regulator-DMGs (TR-DMGs, see Methods). Specifically, we focused on regulators that, when knocked out, lead to decreased expression of the DMGs. Consistent with our expectations, TR-DMGs typically show a positive correlation with each other within wild-type (WT) cells (Binomial test,  $P < 0.00668$ , Fig. 4D).

In scenarios void of an interaction, two transcriptional regulators (R1 and R2) would independently regulate the target gene (G); therefore, a knockout of one regulator should ostensibly not impair the functionality of the other (**Fig. 3.2E**). Constrastingly, should an interaction be present, a knockout of one regulator (e.g., R1) could impact R2’s regulatory capacity over G. This effect could be detected as a change in the gene correlation between R2 and G, when R1 is perturbed (**Fig. 3.2F**). Employing this strategy, we identified 564 genetic interactions amidst 432 unique regulator pairs ( $FDR < 0.1$ , **Fig. 3.2F**). Validating these interactions, analyses incorporating ChIP-seq data from ENCODE<sup>59</sup> show that interacting TR pairs are more likely to have co-localized binding sites proximal to the transcription start site (TSS) of target genes than non-interaction pairs (**Fig. 3.2G**).

As an example, we identified that *IRF1* regulates *LGALS3PB* (evident from differential mean expression analysis) and retains a strong correlation with it in WT cells ( $\rho_{WT} = 0.28$ ). A knockout of *PRDM1* precipitated a significant decrease in the correlation between *IRF1* and *LGALS3PB* ( $\Delta\rho = -0.38$ ), implying a potential interaction between *PRDM1* and *IRF1* in the regulation of *LGALS3PB*. Consistent with these observations, *LGALS3BP* has

binding sites for both *IRF1* and *PRDMB1* in the immediate vicinity of its TSS (**Fig. 3.2H**).

These results demonstrate the capability of correlation analysis via **memento**, especially when applied to forward-genetic screens such as Perturb-seq, to delineate gene sets sharing regulatory elements-albeit participating in diverse pathways-and to reconstruct the genetic interactions of *trans* regulators orchestrating T cell activation.

### 3.3 Genetic analysis of population-scale single-cell

#### RNA-sequencing

The growing availability of scRNA-seq datasets on a population scale has paved the way for mapping genetic variants associated with changes in the expression distribution of proximal genes (*cis*) in specific cell types. Prevailing studies predominantly utilize pseudobulk methods, such as Matrix eQTL, to identify *cis* expression quantitative trait loci (*cis*-eQTLs) impacting mean expression. While linear mixed models have been recently applied to map *cis*-eQTLs in scRNA-seq data, they are hampered by computational inefficiency, a restricted focus on mean comparisons, and susceptibility to misspecification in the underlying parametric model<sup>60</sup>. We posit that, in comparison to pseudobulk methods, **memento**'s superior parameter estimation accuracy and capacity to account for intra- and inter-individual variation will result in increased power to detect *cis*-eQTLs and the discovery of novel variability and correlation QTLs (vQTL and cQTL, respectively). Moreover, the implementation of a highly efficient hierarchical bootstrapping strategy promises applicability to expansive, population-scale scRNA-seq datasets, which could be computationally insurmountable for parametric linear mixed models. To demonstrate, we applied **memento** to reanalyze a pre-existing scRNA-seq dataset, comprising 1.2M PBMCs derived from 160 SLE patients and 90 healthy donors.

The data was analyzed separately for each of the reported cell types: CD4 T cells (T4), CD8

T cells (T8), natural killer cells (NK), classical monocytes (cM), and non-classical monocytes (ncM)<sup>17</sup>. Individuals of East Asian and European ancestries were separately analyzed, with subsequent comparisons enabling a replication analysis between these populations. For every distinct cell type and ancestry group, **memento** mapped *cis* genetic variants-specifically, those within 100kB from the TSS-associated with expression mean, variability, and gene correlation, producing well-calibrated p-values (**Fig. 3.3A**).

A comparative analysis between the power and false positive rate (FPR) of **memento** and Matrix eQTL in detecting *cis*-eQTLs was established against benchmarks provided by the OneK1K study, which comprised of 1000 non-overlapping individuals<sup>18</sup>. Notably in both East Asian and European cohorts, **memento** exhibited higher power in identifying *cis*-eQTLs (AUC=0.85), surpassing Matrix eQTL (AUC=0.81), while maintaining equivalent FPR (**Fig. 3.3A,B**). Overall, **memento** outperformed Matrix eQTL in both populations, replicating 1,606 vs 855 *cis*-eQTLs across cell types in East Asians and, similarly, 1,778 vs 958 in Europeans. Moreover, spanning a range of cohort sizes common for multiplexed scRNA-seq experiments, **memento** achieved an average power gain of 15% for 80 individuals-a metric that increased to 32% for 50 individuals, given an average of 440 cells per individual (**Fig. 3.3C**).

We subsequently explored whether the increased number of *cis*-eQTLs detected by **memento** also improves the enrichment within regions of open chromatin and associations with disease. In the East Asian cohort, *cis*-eQTLs identified by **memento** within specific cell types were more enriched for cell type-specific regions of open chromatin, as annotated by an unrelated study that conducted ATAC-seq on bulk sorted immune cells (p-values for matched cell-types, B  $9.0 \times 10^{-9}$  vs 0.04; T4  $9.3 \times 10^{-4}$  vs 0.11; T8 0.03 vs 0.58; NK  $6.67 \times 10^{-8}$  vs 0.03; cM  $2.1 \times 10^{-11}$  vs 0.67; ncM  $1.0 \times 10^{-6}$  vs 0.46, **Fig. 3.3D,E**). Similar gains in enrichment were observed in the European cohort (**Fig. S3.5**). Further analysis, utilizing LD score regression, found that *cis*-eQTL identified by **memento** also were more enriched for GWAS associations to immune-mediated diseases, thereby suggesting improved fine-mapping



performance (**Fig. S3.6**).

In addition to mapping *cis*-eQTLs, *memento* enables the identification of genetic variants associated with expression variability and gene correlation, offering insights into alternative mechanisms by which genetic variants might influence gene expression. Utilizing *memento*, we identified 10,607 expression variability QTLs (vQTLs) impacting 733 genes across all cell types. For instance, the variability in *HLA-C* expression differed amongst genotypes of 6:31326612 (**Fig. 3.3F**), with the A allele amplifying the expression variability of *HLA-C* without notably affecting its mean (**Fig. 3.3G**). For mapping correlation QTLs (cQTLs), we focused on testing the correlation between genes possessing at least one significant *cis*-eQTL and known transcription factors, thereby specifically testing the hypothesis that genetic variants might modulate the effect of transcription factors on gene expression. We mapped 3,726 cQTLs for 238 gene pairs across all cell types. For example, the SNP at 12:69688073 is associated not only with the mean expression of *LYZ*, but also the correlation between *JUNB* and *LYZ*. Intriguingly, a *JUNB* binding site exists within 1kbp of the SNP, suggesting that *JUNB* may serve as a *trans* regulator for *LYZ*, with the regulatory strength being influenced by the genotype at this SNP.

These findings underscore *memento* as a scalable approach for genetic analyses of population-scale scRNA-seq data, delivering higher statistical power for identifying *cis*-eQTLs and introducing the capability for mapping vQTLs and cQTLs. These advances not only improve the fine mapping of disease associations but also unveils novel mechanisms whereby genetic variants may modulate gene expression.

### 3.4 Census-scale differential expression analysis across cell types, individuals, and disease states

The above applications showcased the broad applicability of `memento` for generalized differential expression analysis across a diverse set of individual datasets, including the analysis of the temporal response of tracheal epithelial cells stimulated by IFN, the mapping of gene regulatory networks from Perturb-seq data of CD4+ T cells, and large-scale genetic analysis of gene expression across single cells. Through these applications and simulations, we showed that `memento` consistently outperforms existing methods, delivers a unique feature set to compare variances and covariation in addition to the mean, and is extremely efficient enabling scaling to a million cells and tens of replicates.

The emergence of massive repositories of single-cell data across the world has raised novel needs for computational techniques that can efficiently compare datasets while ensuring properly calibrated statistical behavior. As of November 2023, CELLxGENE Discover includes 50 million unique cells across 1,102 datasets, with over thousands of individuals represented, with its Census API providing access to most of these data<sup>61</sup>. Unlike a scRNA-seq dataset generated by a single research project with a focused hypothesis, users of CELLxGENE Discover access this resource with a diverse array of comparative analyses in mind. For example, one user may be interested in differences in expression between the same cell types residing in different organ systems. Another user may be interested in differences in expression between the same cell types across individuals with different disease status. In any scRNA-seq dataset with labeled cell types, there is a large number of possible comparisons between cell groups (**Fig. 3.4AB**). Furthermore, multiple datasets may be combined to improve the power of comparisons between the same cell groups that exist across datasets.

Differential expression methods powering queries within the census need to efficiently perform accurate, well-calibrated comparisons between user-defined cell groups across

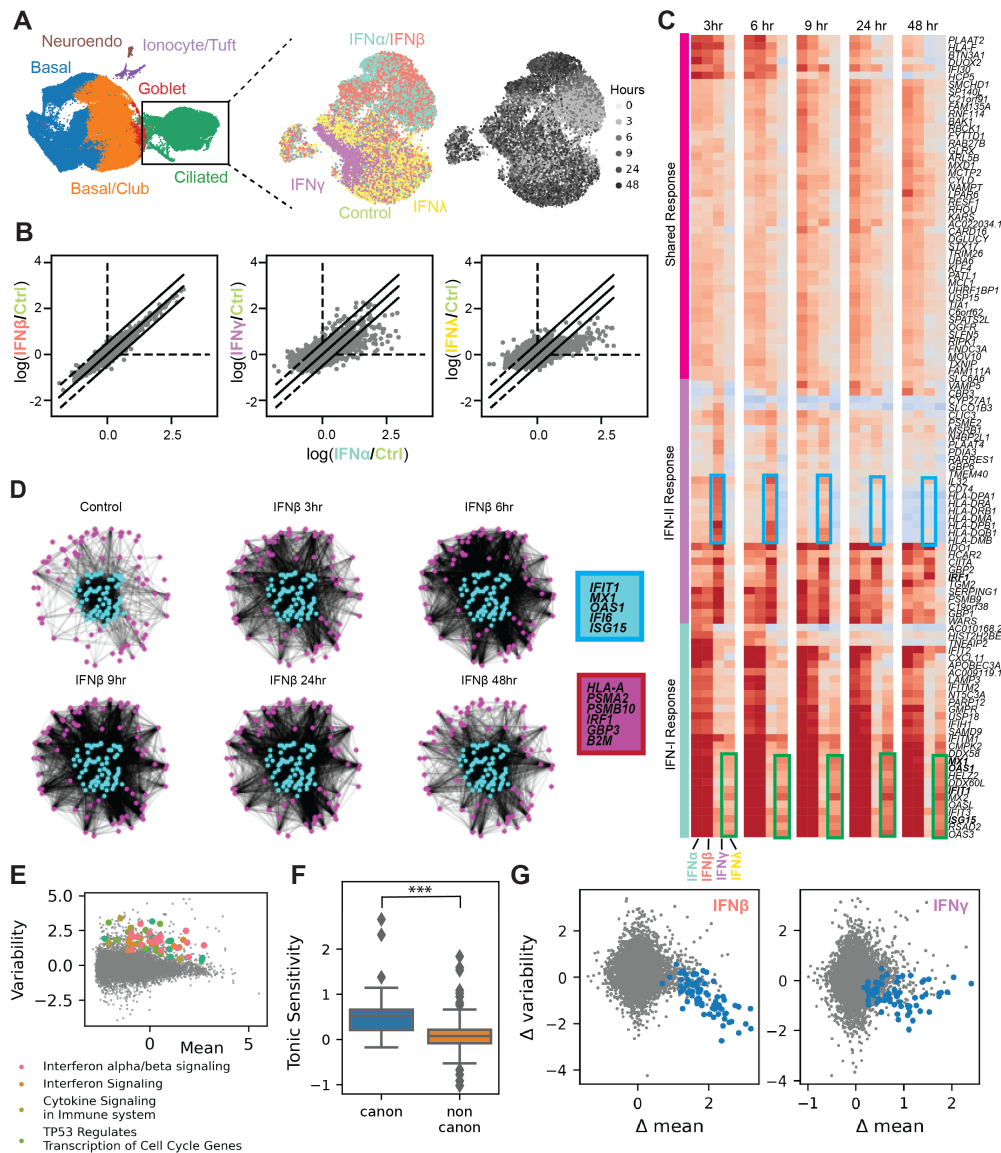
datasets, delivering results near real-time speed for web portal integration. Although **memento** demonstrates excellent scalability with increasing cell numbers, as shown in **Fig. 2.2F**, its real-time result delivery is constrained by the necessity of performing bootstrap operations for each comparison, a limitation that becomes more pronounced when subsets contain multiple biological and technical replicates. To extend the broad applicability of **memento**, we collaborated with the CZI to utilize the CELLxGENE Discover Census API to perform bootstrap operations and quantify uncertainty for predefined cell groups across the entire corpus (see Methods). This extension allows for the precomputation of standard errors, which are then utilized to enable near real-time differential expression analysis via weighted least squares. Consequently, the standard errors derived from this precomputed mode provide an effective approximation of the bootstrap method employed in the full mode, streamlining the analysis process.

To evaluate the agreement between **memento** in its precomputed mode and the full mode, we conducted a differential expression analysis comparing CD4 T cells and classical monocytes from a single donor in the lupus dataset (referenced in **Fig. 3.3**), also included in the CELLxGENE Discover. Given that the analysis involved the same underlying data, we anticipated highly similar results. The primary difference would be attributed to the two **memento** versions, with the precomputed mode utilizing estimated cell sizes from the entire CELLxGENE Discover dataset. This anticipation was confirmed by observing a robust correlation in the effect size estimates (**Fig. S3.7**) between the full and approximate, precomputed modes. A similarly strong correlation was noted in the significance levels, indicated by  $-\log_{10}(P - value)$  (**Fig. 3.4CD**). Remarkably, the computation time for determining effect size and P-value was significantly reduced compared to executing **memento** in full mode for various cell group comparisons (**Fig. 3.4E**).

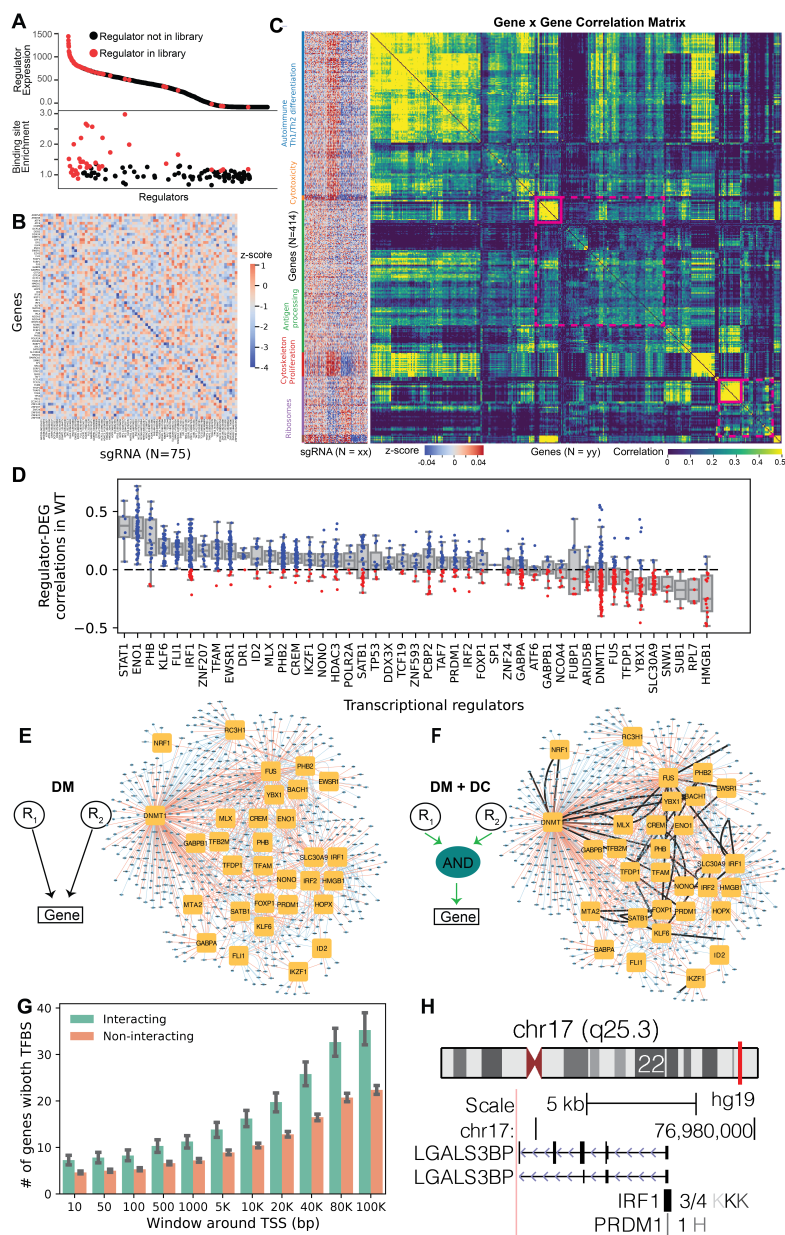
A unique application of **memento** on large-scale census data lies in its improved power to compare cell groups, particularly beneficial for those that are rare in individual datasets. To

illustrate this, we utilized `memento` in its precomputed mode to identify DM genes between conventional and plasmacytoid dendritic cells. These cell types constitute 5.8% and 4.0%, respectively, of the scRNA-seq datasets of immune cells within the CELLxGENE Discover (**Fig. 3.4F**). In analyzing 23 separate datasets in the CELLxGENE Discover, encompassing 362,619 total cells, we found that a joint analysis across these datasets significantly increased the statistical power compared to analyses of any single dataset (**Fig. 3.4G**). These results underscore that the efficiency of `memento`'s moment estimators and the adaptability of its bootstrap approach enable its effective application in expansive census repositories.

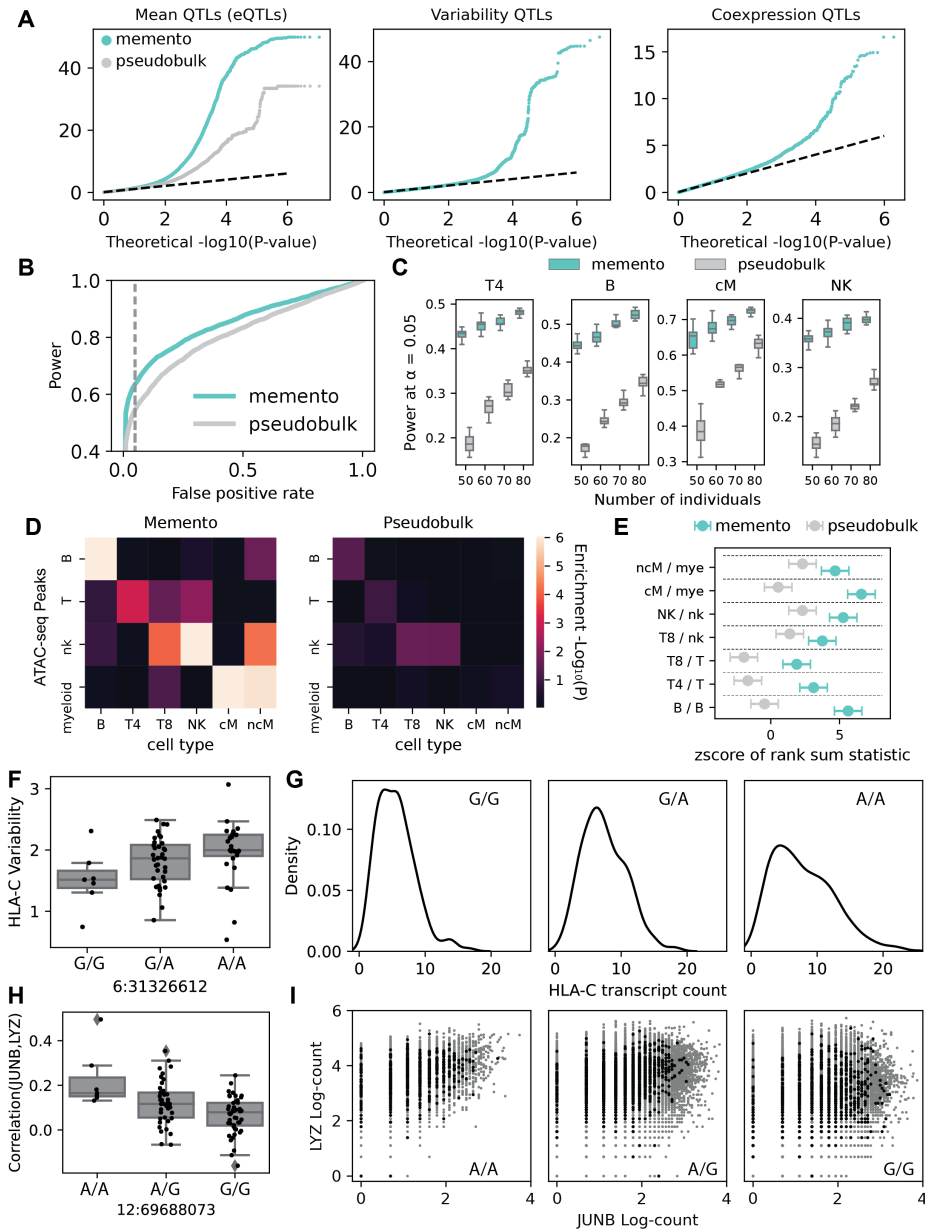
## 3.5 Figures



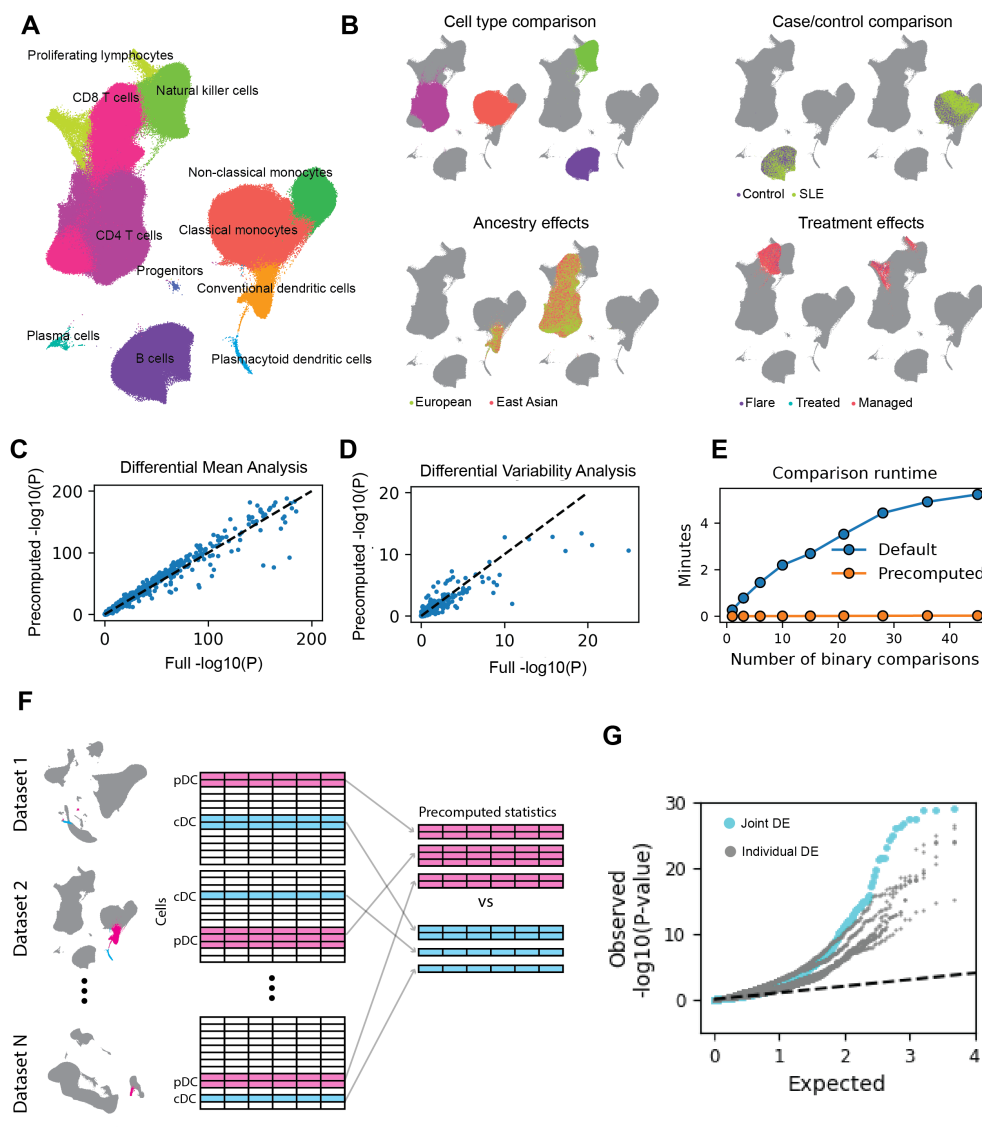
**Figure 3.1: Mapping transcriptional response of human bronchial epithelial cells to extracellular interferon using memento.** (A) UMAPs of the entire HBEC dataset colored by identified cell types (left), zoomed in ciliated cells colored by stimulation (center), and time labels (right). (B) Log fold-change (LFC) of mean expression in response to IFN- $\alpha$  (x-axis) against LFC in response to IFN- $\beta$  (left), IFN- $\gamma$  (middle), and IFN- $\lambda$  (right) after 6 hours. (C) Hierarchically clustered heatmaps of LFC in response to the four types of interferons (columns within each heatmap) across 5 timepoints. (D) Gene coexpression network over time where magenta nodes depict canonical ISGs and cyan nodes depict noncanonical ISGs. (E) Baseline expression variability (y-axis) versus mean (x-axis) in ciliated cells. (F) Tonic sensitivity (y-axis) for canonical and non-canonical ISGs (x-axis). \*\*\* indicates  $P < 0.001$ . (G) Change in variability (y-axis) against the change in the mean (x-axis) in response to IFN- $\beta$  (left) and IFN- $\gamma$  (right). Blue dots represent canonical ISGs.



**Figure 3.2: Reconstructing gene regulatory networks of T cell activation using Perturb-seq and memento.** (A) Selection criteria for perturbed regulators in this study. (B) Heatmap of average gene expression for each gene across cells perturbed by the corresponding sgRNA. (C) Left: Heatmap of average gene expression for DMGs across cells perturbed by the corresponding sgRNA. Right: Gene-gene correlation matrix for the same DMGs estimated from WT cells. (D) Correlation between each regulator and its downstream genes in WT cells. (E) Bipartite gene regulatory network that do not account for interaction between regulators. (F) Gene regulatory network including genetic interactions between regulators. (G) Number of genes with binding sites for pairs of interacting or non-interacting regulators across varying windows of the TSS. (H) Chromosomal location of *LGALS3BP* and binding sites for IRF1 and PRDM1, predicted to interact using DM and DC analysis.



**Figure 3.3: Mapping of mean QTL, variability QTL, and correlation QTL using memento.** (A) Quantile-quantile (QQ) plots for expected p-values (y-axis) computed by memento versus theoretical p-values (x-axis). For mean QTLs, QQ-plot of p-values from pseudobulk approach (Matrix eQTL) is overlaid. (B) ROC curve for recovery of mean QTLs. (C) Power of eQTL recovery (y-axis) of memento and pseudobulk method across different numbers of individuals. (D) Enrichment of cell-type specific eQTLs in cell-type specific ATAC-peaks. (E) Enrichment of eQTLs detected in each cell type for cell-type-specific ATAC-peaks. (F) An example of a variability QTL. (G) Histogram showing distribution of *HLA-C* expression for a representative individual of each genotype. (H) An example of a correlation QTL. (I) Scatterplot of expression of *LYZ* (y-axis) against the expression of *JUNB* (x-axis) across single cells from all donors (grey) and a representative individual (black).



**Figure 3.4: Extending memento for near real-time differential expression analysis within CZI CELLxGENE Discover.** (A) UMAP of the SLE PBMC dataset within CELLxGENE. (B) Enumeration of different comparisons that can be made within and between groups of cells. Comparisons of significance (P-value) between the precomputed and full modes for (C) differential mean and (D) differential variability analyses. (E) Runtime as a function of number of comparisons made, at query time (excluding precomputation). (F) Schematic of multiple datasets analyzed with CELLxGENE identifying DMGs between pDCs and cDCs. (G) QQ-plot of P-values from comparing pDCs and cDCs combining many datasets (cyan) and using each dataset alone (grey).



## 3.6 Methods

### HTEC interferon stimulation experiment

Human tracheal epithelial cells were harvested from deceased organ donors according to established protocols (PMID: 1616056). Frozen cell aliquots were reactivated and cultured in epithelial growth media (EGM) [3:1 (v/v) F-12 Nutrient Mixture (Gibco)–Dulbecco’s modified Eagle’s medium (Invitrogen), 5% fetal bovine serum (Gibco), 0.4 ug/mL hydrocortisone (Sigma-Aldrich), 5 ug/mL insulin (Sigma-Aldrich), 8.4 ng/mL cholera toxin (Sigma-Aldrich), 10 ng/mL epidermal growth factor (Invitrogen), 24 ug/mL adenine (Sigma-Aldrich), and 10 uM Y-27632 (Enzo Life Sciences)] on 10 mm dishes coated with rat tail collagen (Sigma-Aldrich). EGM was changed three times a week until dishes were confluent, at which point the cells were passaged with 0.25% trypsin for 30 minutes. For air liquid interface culture, expanded basal cells were plated at 50,000 cells per 6.5 mm transwell insert (Corning 3470) coated with human placental collagen (Sigma-Aldrich) and cultured with Pneumacult ALI (StemCell) for 21-28 days according to the manufacturer’s instructions. Starting on day 27, interferon stimulation (IFN- $\beta$ : 10 ng/ml, IFN- $\alpha$ 2: 10 ng/ml, IFN- $\gamma$ : 10 ng/ml, IFN- $\lambda$ 2: 10 ng/ml) was added at hours 0, 24, 39, 42, and 45 prior to harvesting (For final timepoints 3, 6, 9, 24, and 48 hours). On the day of harvest, basal media was aspirated and both basal and apical chambers were rinsed twice with PBS. Following two washes, trypsin-EDTA (0.25% Fisher cat. 25200072) was added to both the basal and apical chambers (300 ul basal, 100 ul apical) and incubated for 30 minutes at 37°C while pipette mixing every 10 minutes. Trypsinization was quenched with 300 ul of maintenance media and transferred to a 1.5ml eppendorf tube (eppendorf cat. 022431021) and centrifuged at 350xg for 5 minutes at 4°C. Cells were resuspended in 94 ul of cell staining buffer (Biolegend cat. 420201) and blocked with 5 ul of TruStain FcX (Biolegend cat. 422302) for 10 minutes on ice. Blocked cells were stained with 1 ul of Biolegend Totalseq-B hashtags (Biolegend Totalseq-B hashtags 1-11) for 30 minutes on ice. Staining was quenched with 1 ml of cell staining buffer

and spun at 300xg for 5 minutes at 4°C prior to two more washes with 1 ml of cell staining buffer. Cells were resuspended in 100 ul of 0.05% BSA in PBS and counted via Countess II (Fisher cat. A27977). Counted cells were pooled equally into two pools and spun at 300xg for 5 minutes at 4°C. Cells were strained through a 100  $\mu$ M filter (Corning cat. 431752) prior to a final count and each pool was loaded onto two 10x 3'v3 lanes. Libraries were prepared as described in the 10x 3'v3 user guide. Samples were sequenced on three lanes of NovaSeq S4.

## Clustering the HTEC transcriptomes

We performed filtering, normalization, and clustering with the scanpy<sup>32</sup> suite of tools using the default values. Cell types were manually identified based on previously known marker genes for HTECs<sup>52</sup>.

Similar to the rIFNB1 dataset, we selected genes where the mean observed expression  $\mathbb{E}[Y_{cg}] = 0.07$ , which was the reliability limit for this experiment.

## Clustering the correlation matrices for genes with differential mean expression

DMGs in ciliated cells were identified by using `memento` by comparing each stimulation and timepoint to the unstimulated control. The correlation between the DMGs were computed using `memento` for each timepoint in IFN- $\beta$  stimulation condition. This correlation matrix at timepoint 6hr was then clustered using the `AgglomerativeClustering` function in `sklearn` python package. Top 4 clusters in terms of gene number were chosen for plotting.

## Identifying highly variable genes at baseline

We used `memento` in the one-sample mode to compute the donor-averaged expression mean and variability for each gene in the transcriptome that had greater than 0.07 mean UMI count. We then performed gene set enrichment analysis using `EnrichR` to get the

significantly enriched gene sets.

## **Study subjects and genotyping for Perturb-seq**

Our samples were enrolled in PhenoGenetic study (age 18 to 56, average 29.9), as part of the Immvar cohort(20), which were recruited in the Greater Boston Area. Each donor gave written consent to participate and were healthy, without any history of inflammatory disease, autoimmune disease, chronic metabolic disorders or chronic infectious disorders. We genotyped 56 caucasian samples on the OmniExpressExome54 chip, and excluded 2080 SNPs with a call rate  $< 90\%$  (0.22% of total), 1521 SNPs with Hardy Weinberg  $P < 0.0001$  (0.16%) and 259,860 SNPs with MAF  $< 0.1$  (27.04%) out of the total 960,919 SNPs profiled. The Michigan Imputation Server was used to impute these genotypes with the Haplotype Reference Consortium Panel Version r1.1. After genotype imputation had 5,324,560 SNPs, which were then subsetted for our nine donors.

## **Regulator target identification and CROP-seq library generation**

Our library contained targeted 140 regulators (transcription factors and RNA-binding proteins) with 2 sgRNAs each. Each regulator was unbiasedly chosen using gene expression and accessibility data from activated CD4+ T cells in 95 and 105 healthy donors(18). To get the highly expressed regulators using RNA-seq data, we performed a TMM normalization and took the upper quartile of highly expressed genes and subsetted those that were regulators. To get the regulators with highly accessible binding sites using ATAC-seq data, we enriched for all binding sites on the HOMER database(71) in activated accessible chromatin regions. We took the union of the highly expressed regulators and accessible binding sites, for a total of 140 regulators (Fig. 1B).

The backbone plasmid used to clone the CROP-Seq library was CROPseq-Guide-Puro(28), purchased from Addgene (Addgene. Plasmid #86708). We used two sgRNAs oligo sequences

from the Brunello library<sup>(88)</sup> for each of our chosen 140 regulators. Oligos for the sgRNA library were purchased from Integrated DNA Technologies (IDT) and cloned into the CROPseq plasmid backbone using the methods described by Datlinger et al.<sup>13</sup>. Lentivirus was produced using the UCSF ViraCore.

## **SLICE experiment and sequencing**

Primary human CD4<sup>+</sup> T cells were isolated from peripheral blood mononuclear cells (PBMCs) by magnetic negative selection using the EasySep Human CD4<sup>+</sup> T Cell Isolation Kit (STEMCELL, Cat #17952). Cells were cultured in X-Vivo media, consisting of X-Vivo15 medium (Lonza, Cat #04-418Q) with 5% Fetal Calf Serum, 50mM 2-mercaptoethanol, and 10mM N-Acetyl L-Cysteine. On the day of isolation (Day 1), cells were rested in media without stimulation for 24 hours. The day after isolation (Day 2), cells were stimulated with ImmunoCult Human CD3/CD28 T Cell Activator (STEMCELL, Cat #10971) and IL-2 at 50U/mL. 24 hours post stimulation (Day 3), 1 uL of lentivirus was added directly to cultured T cells and gently mixed. Following 24 hours (Day 4), cells were collected, pelleted, and washed in PBS twice. Then, cells were resuspended in Lonza electroporation buffer P3 (Lonza, Cat #V4XP-3032). Cas9 protein (MacroLab, Berkeley, 40mM stock) was added to the cell suspension at a 1:10 v/v ratio. Cells were transferred to a 96 well electroporation cuvette plate (Lonza, cat #VVPA-1002) for nucleofection using the Lonza Nucleofector 96-well Shuttle System and pulse code EH115 (Lonza, cat #VVPA-1002). Immediately after electroporation, pre-warmed media was added to each electroporation well, and 96-well plate was placed at 37 degrees for 20 minutes. Cells were then transferred to culture vessels in X-Vivo media containing 50U/mL IL-2 at 1e6 cells /mL in appropriate tissue culture vessels. Two days later, 1.5ug/mL Puromycin was added in culture media for selection. Cells were expanded every two days, adding fresh media with IL-2 at 50U/mL. Cells were maintained at a cell density of 1e6 cells /mL. On the final day (Day 13) of the experiment, cells from each of the nine donors were counted using Vi-CELL XR and pooled

at equal numbers to obtain a final 180,000 cells in 60 uL of PBS. The pooled cells were then processed by UCSF Institute for Human Genetics (IHG) Genomics Core using 16 wells of 10X Chromium Single Cell v2 (PN-120237), as per manufacturer's protocol, with each well being separately indexed. The final library was sequenced on two lanes on the Nova-seq for a total of 6.7B reads. To maximize the probability of detecting sgRNAs in cells, we further amplified and sequenced the sgRNA transcripts from the 10X cDNA library to near saturation as previously described<sup>62</sup> (98%).

## Estimation of cutting efficiency

The cutting efficiencies are estimated as the proportion of DNA in bulk that contained a specific indel (by readcount) normalized by the relative proportion of cells with a specific sgRNA found in the experiment.

It is important to note that while neither the denominator or the numerator cannot be a number larger than 1, our estimate of the proportion of cells with a specific guide used for normalization contains error, leading to a handful of guides with a cutting efficiency  $> 1$ .

## Visualizing gene regulatory networks

To generate the GRNs in 3.2E, we first used a list of pairs of regulator to their differential-mean expressed genes to define a bipartite graph, which was then visualized in Cytoscape. We then added the connections between the interacting pairs of regulators discovered by differentially correlated genes (DCGs) in the same previously visualized network (3.2F).

## Identifying candidate interactions for differential correlation analysis

For a transcriptional regulator TR, we first identified all of the DMGs where the TR acts as a transcriptional activator, with the DM coefficient less than 0 across the KO. We then

computed the correlation between each TR-DMG pair in WT cells, and constructed the final set of TR-DMG pairs by selected those that had a significant correlation in WT ( $\rho > 0.1$ ). For each of these TR-DMG pairs, we tested for differential correlation across various sgRNAs targeting transcriptional regulators other than TR. The final set of interactions were called by filtering for  $FDR < 0.1$ .

## **Counting genes with shared TFBS for pairs of transcription factors**

For a pair of transcriptional factors TF1 and TF2, we first identified their transcription binding sites (TFBSs) during the ChIP-seq data in the ENCODE datasets. We then took the locations of known gene transcriptional start sites (TSSs) and measured the distance of the nearest TFBS for each TF for each TSS. We then counted the number of genes that have TFBSs of both TF1 and TF2 within a series of window sizes near the TSS, ranging from 10 base pairs to 100K basepairs. We performed this procedure for pairs of TFs chosen at random and also pairs of TFs identified as interacting using differential correlation analysis.

## **Assessing the tonic sensitivity of ISGs**

We used tonic sensitivity measurements from Gough et al. where the authors compared the expression of ISGs in IFNAR1-KO and WT macrophages<sup>53</sup>. The fold-change between those two groups were defined as the tonic sensitivity, which is the number we use in Fig. 3.1D.

## **eQTL discovery using pseudobulk approach and memento**

We used the single cell dataset generated by Perez et al. that profiled peripheral blood mononuclear cells in individuals with systemic lupus erythematosus (SLE) and healthy controls. We maintained the same cell type classifications used in that study.

To identify eQTLs using the pseudobulk approach, we first created pseudobulks at the cell-type and individual level by normalizing each cell expression with total UMI count per cell, taking

the average for each gene across all individuals, and computing  $\log(x + 1)$  for each mean. We filtered genes that had a lower than 0.01 mean UMI counts in the single cell dataset.

We ran Matrix eQTL for each of the Asian and European populations separately, using the same set of genotypes and covariates used by Perez et al.<sup>17</sup>. For `memento`, we also performed the test separately for the two populations, using the same genotypes and covariates. We used the hierarchical resampling mode for `memento`.

### **Enrichment of eQTLs in ATAC peaks**

We used the same set of ATAC peaks used by Perez et al.<sup>17</sup>. For each SNP, we labeled whether that a cell type specific ATAC peak covered the location of the SNP. We then compared the p-values of the eQTL candidates in a cell-type peak to those of the candidates outside of ATAC peaks using the Wilcoxon Rank Sum test.

### **Comparison of eQTLs with OneK1K cohort**

To compute the ROC curve and perform power analysis in 3.3, we compared the eQTLs we discovered using the two approaches to the eQTLs reported by Yazar et al.<sup>18</sup>. We used this much larger dataset as the gold standard to compare methodologies applied to the SLE dataset. Specifically, we calculated power as the proportion of Onek1k hits we were able to replicate with the smaller dataset. We calculated false positive rate by shuffling the genotypes of individuals (while keeping individual-cell assignments intact) and calculating the proportion of SNP-gene pairs with  $P < 0.05$ .

### **Precomputation of estimates and standard errors in the CELLxGENE Discover database**

The CELLxGENE Census database provides RNA expression counts as an  $M \times N$  matrix comprised of  $M$  cells and  $N$  genes. Each cell is annotated with metadata specifying its cell

type, dataset, tissue, assay type, donor, disease, sex, development stage, ethnicity and suspension type. The Census data is sparse, in that if the measured expression of a given gene for a given cell is not positive, it is not explicitly stored.

From these Census data, we grouped the cells by their annotation values for cell type, dataset, tissue, assay type, donor, disease, sex, development stage, ethnicity and suspension type. Then, for every group of cells and every expressed gene within a given cell group, we computed the logs of the mean, standard error of the mean, variance, and standard error of the variance using the same estimator and resampling strategy outlined above.

These precomputed values are then saved so that repeated differential expression analyses can be efficiently performed without recomputing these estimators. A Census data comprising 30M cells is reduced to 140K cell groups, and is therefore 2 orders of magnitude smaller in size.

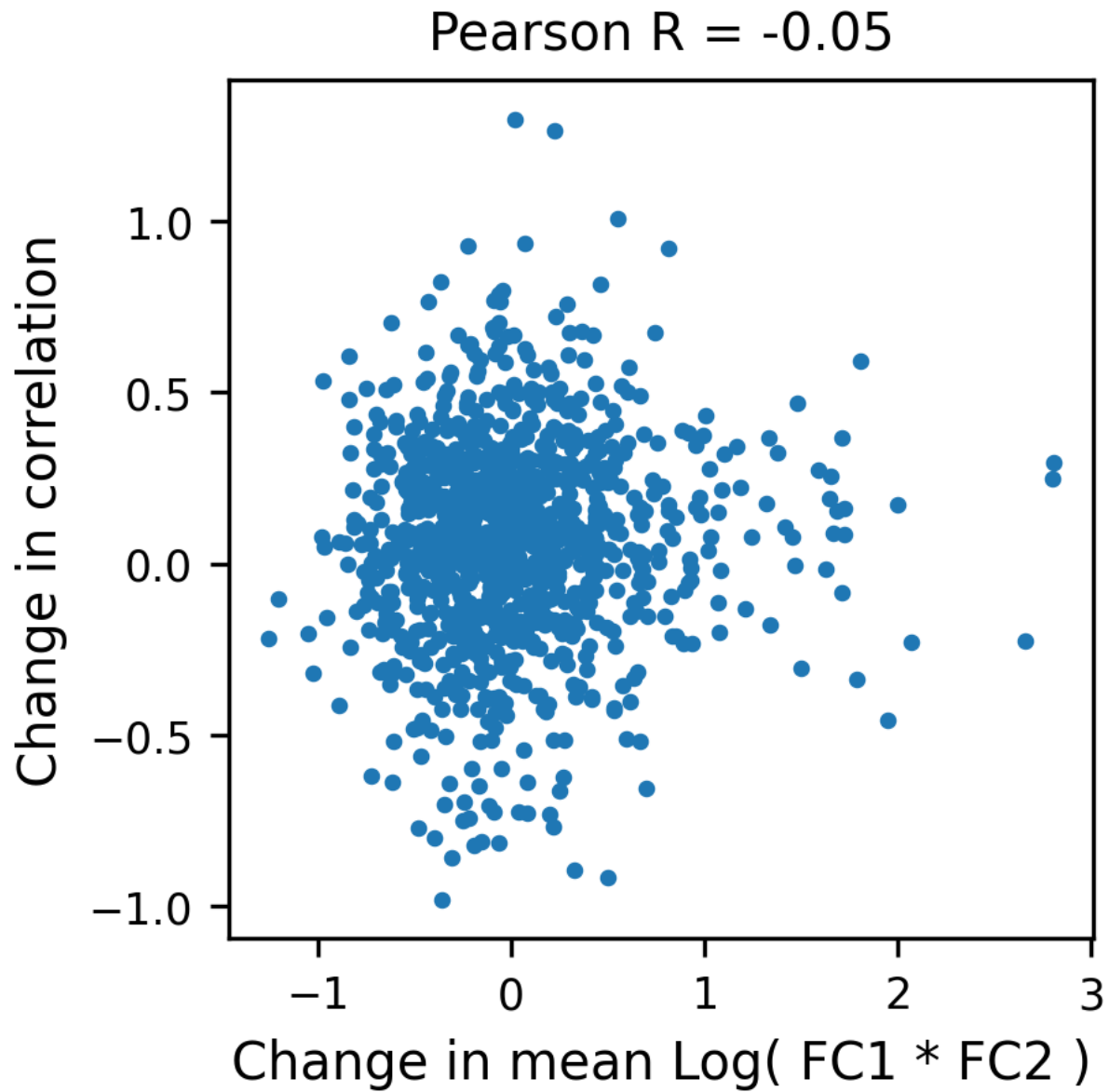
## **Hypothesis testing using precomputed standard errors**

To compute differential expression between two distinct groups of cells that differ by a specified treatment, two subsets of the precomputed data can be retrieved by filtering the precomputed data by two distinct values of a specified cell annotation. All of the remaining cell annotations are then treated as covariates when computing differential expression.

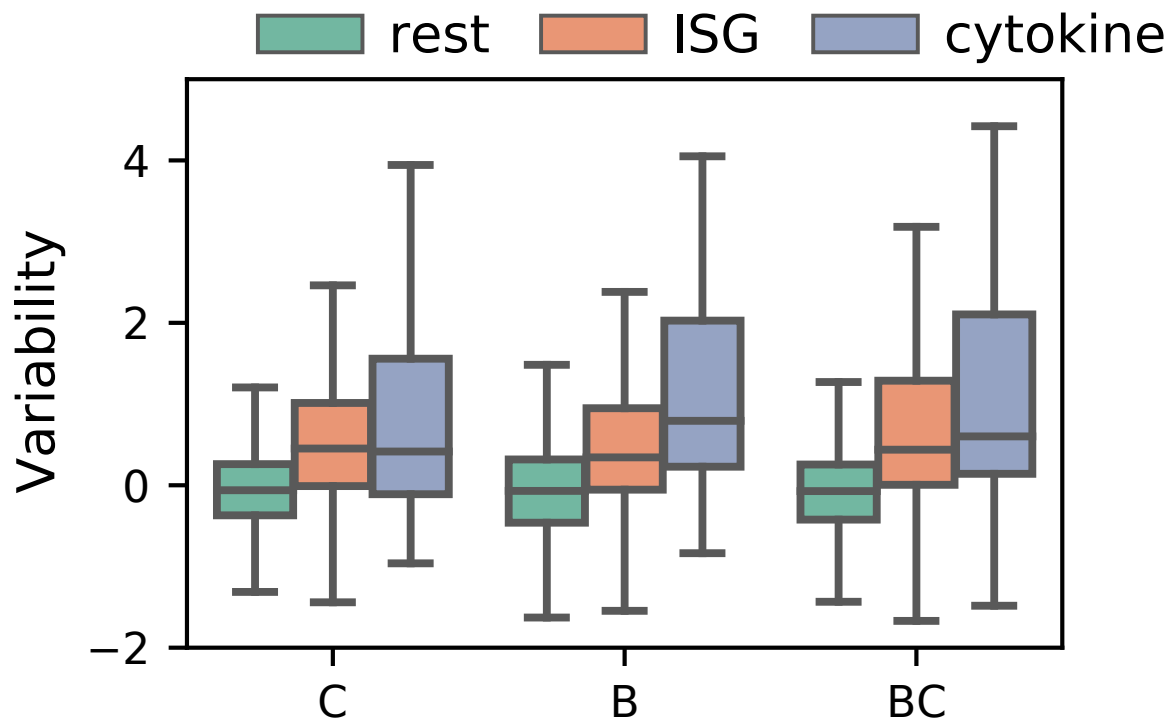
For the cell type comparisons presented in the paper, we used basic weighted least squares (WLS) to incorporate the precomputed mean and residual variance estimates (as response variable) along with their standard errors (as weights). To perform DE across all 23 datasets in (**Figure 3.4**), we used the donor covariates as one-hot encoded variables as well as their interaction terms with the cell type one-hot encoded variable.



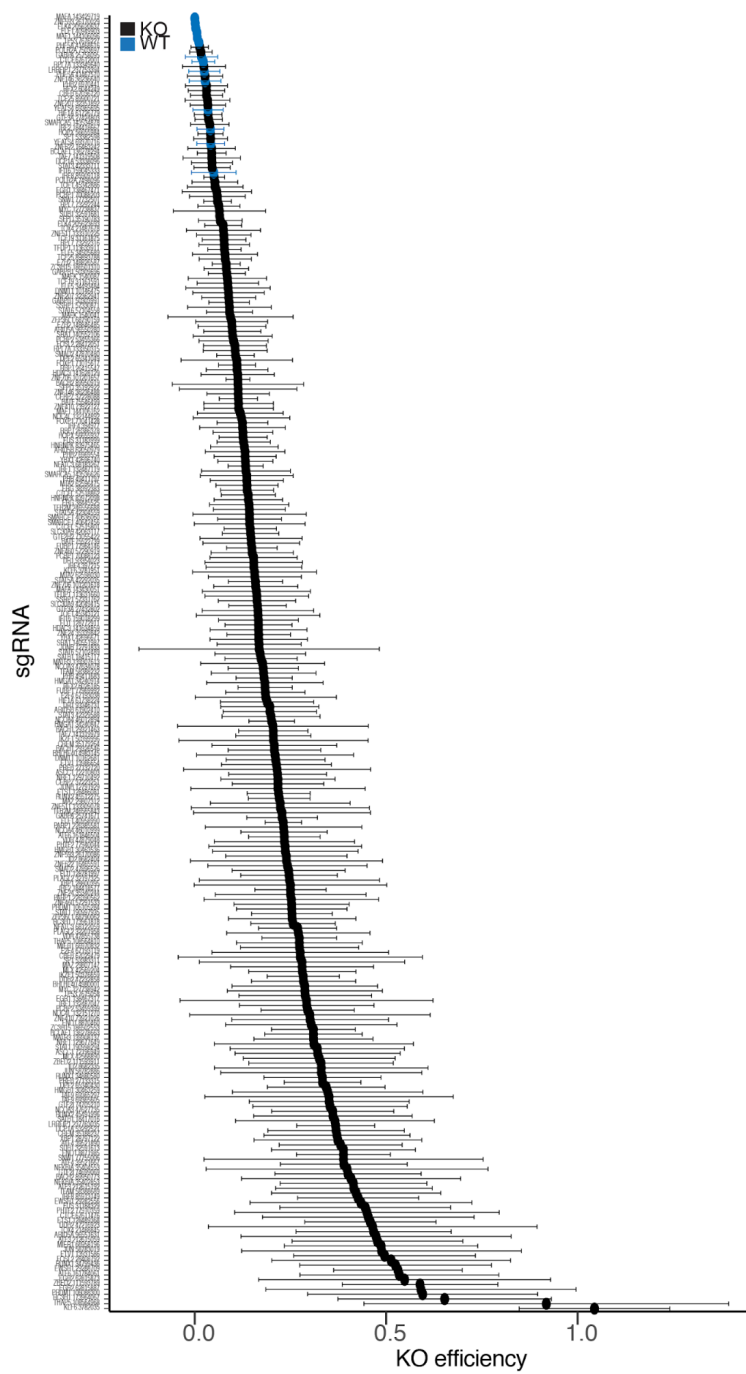
### 3.7 Supplementary figures



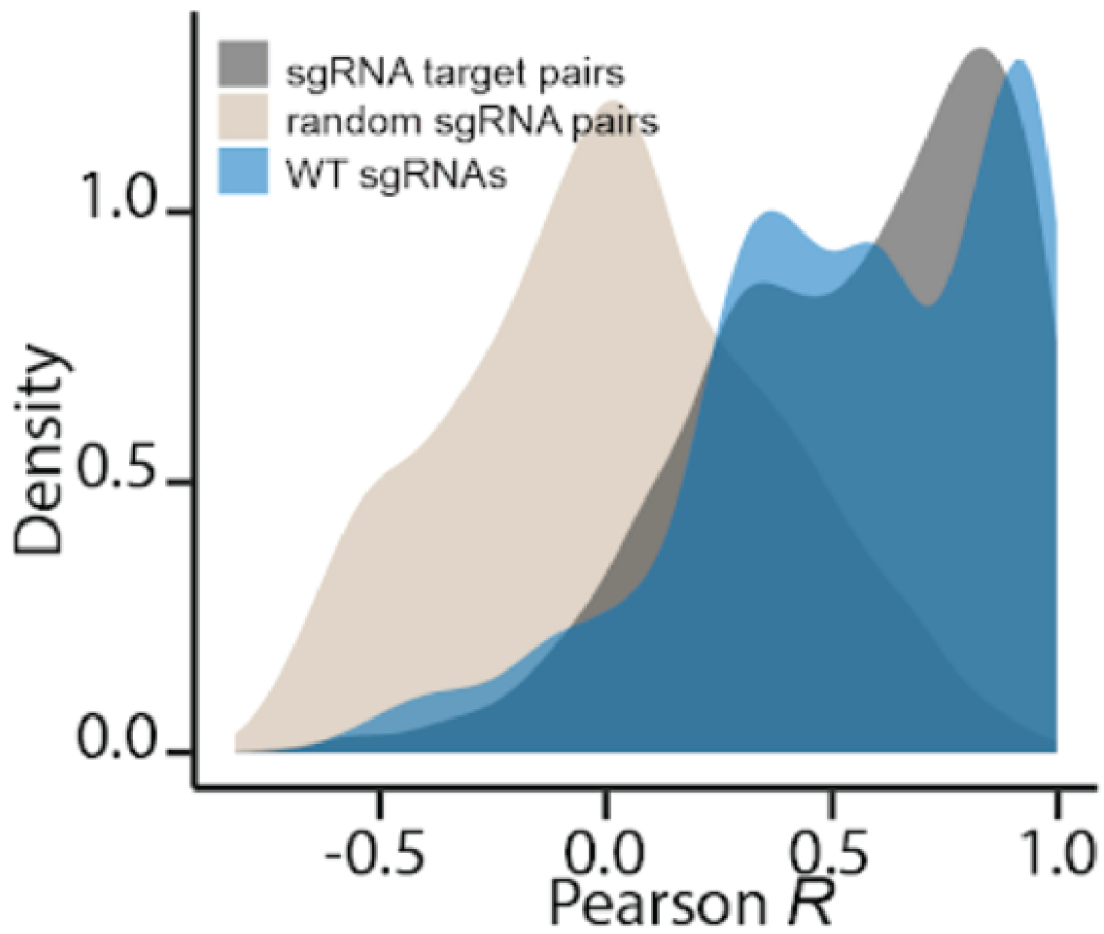
**Figure S3.1:** Change in correlation between two genes (y-axis) vs the product of the changes in mean (x-axis).



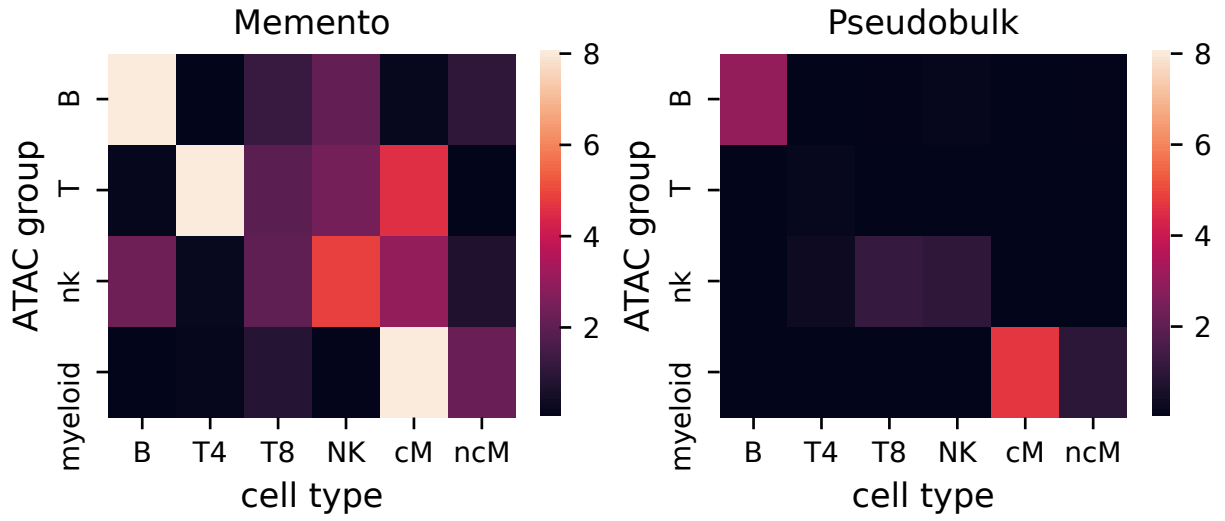
**Figure S3.2:** Gene expression variability (y-axis) for each class of genes across cell types.



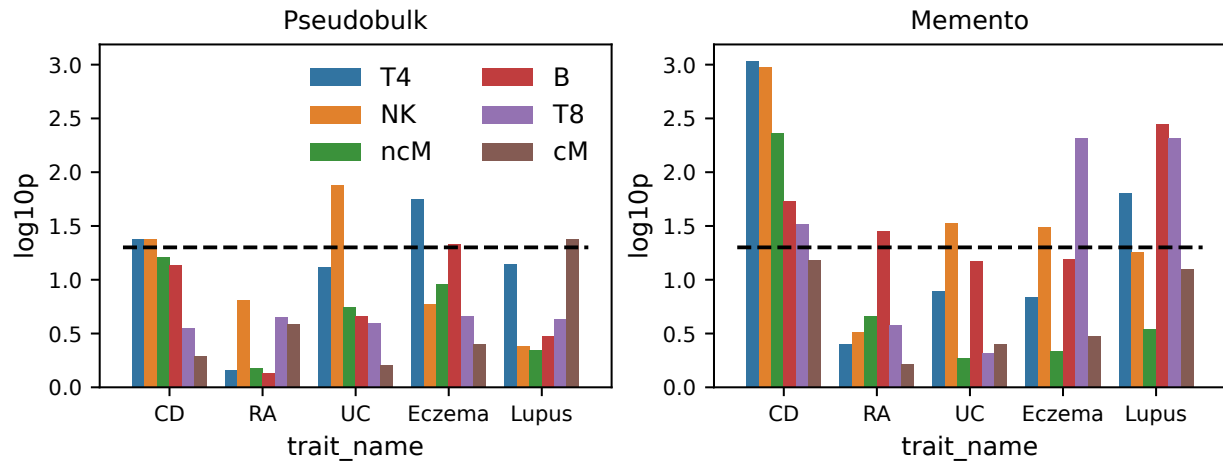
**Figure S3.3:** We estimated the average sgRNA KO efficiency (x-axis) per sgRNA (y-axis). Each point represents the average KO efficiency and error bars are the standard deviations across donors.



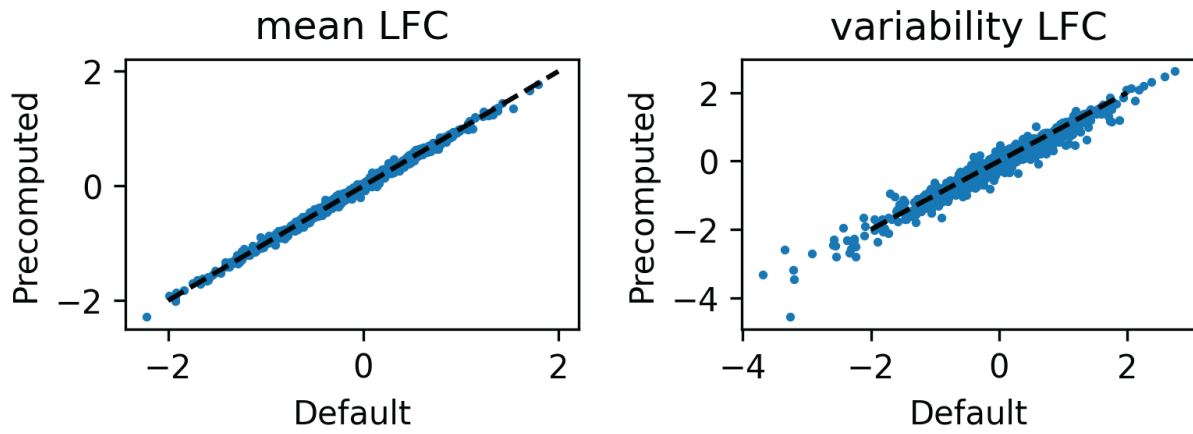
**Figure S3.4:** Distribution of transcriptome correlations for WT guides, guides targeting the same gene, and random pairs of guides.



**Figure S3.5:** Enrichment of eQTLs in cell-type specific ATAC peaks (Europeans).



**Figure S3.6:** LDSC-score regression enrichment for diseases using eGenes found via *memento* and the pseudobulk method.



**Figure S3.7:** Comparison of effect sizes computed using full and precomputed versions of memento

# Chapter 4

## Discussion and conclusion

Fueled by the development of scalable workflows, there is an emergence of scRNA-seq datasets where the quantitative comparison of gene expression distributions between groups of cells is a critical task. These include endeavors to compare single-cell expression profiles between experimental conditions<sup>12</sup>, disparate genetic perturbations induced by genome editing<sup>14,63</sup>, and individuals inheriting different alleles<sup>16–18</sup>. Initial observations that experimental and genetic perturbations predominantly induce subtle shifts in gene expression rather than unequivocal cell states have necessitated the need for methods adept at comparing gene expression distributions. However, scalable computational methods that facilitate hypothesis testing over large numbers of cells and an extensive array of covariates (e.g. hundreds of *in vitro* perturbations or millions of genetic polymorphisms) are still scarce. Moreover, even fewer methods currently test for differences in the variability of gene expression and gene correlations, unique parameters captured by single cell RNA-sequencing.

Here, we introduced **memento**, an end-to-end method for the quantitative analysis of scRNA-seq data theoretically scalable to millions of cells. **memento** is developed with two pivotal innovations: method of moments estimators modeling scRNA-seq via a hypergeometric sampling process and an efficient bootstrapping strategy to construct precise



confidence intervals around parameter estimates, exploiting the sparsity of scRNA-seq data. The utilization of method of moments estimators imparts twofold advantages over other approaches. First, our approach delineates biological and technical sources of noise, enabling the accurate characterization of biological variation. This feature of **memento** addresses recent calls for hierarchical parametric modeling of the measurement noise of scRNA-seq while only considering biological variation for estimation and inference<sup>22</sup>. Second, our approach circumvents the need to repetitively compute overall likelihood, enabling instantaneous computation of the pertinent parameters. The multinomial approximation of hypergeometric sampling has been used to theoretically derive the baseline noise in scRNA-seq<sup>34</sup> and to design dimensionality reduction techniques for count data<sup>64</sup>. The Poisson approximation of the binomial (which in turn approximates the hypergeometric), has been used to derive empirical Bayes estimators to inform the optimal design of scRNA-seq experiments<sup>37</sup>. While our estimators are derived focusing on scRNA-seq workflows where cell-to-cell differences in transcript sampling frequencies  $q$  are small, the hypergeometric formulation is amenable to models where  $q$  varies significantly between compartments (e.g. sci-rna-seq<sup>30</sup>), provided that  $N_c$  and  $q$  can be estimated separately. Because of the modular and flexible nature of **memento**, we further anticipate that our modeling framework could be extended to alternative scRNA-seq workflows that use hybridization instead of reverse transcription<sup>65</sup> and spatial transcriptomics data<sup>66</sup>. Analyses of emerging multimodal workflows (e.g., ATAC-seq and CITE-seq) should also be possible by modifying the method-of-moments estimators to correctly capture sources of technical variation unique to each assay.

The implementation of method of moments estimators for hierarchical models universally contends with the challenge of establishing confidence intervals via resampling, given that incorporating the sampling process into deriving analytical confidence intervals and p-values can materialize as exceedingly complex without further assumptions. Although resampling can be computationally prohibitive, particularly when cell numbers are large, our

employment of the approximate bootstrap resamples the number of unique counts as opposed to the number of single cells. In various cell count subsamples, the number of unique counts increased sub-linearly with the number of cells (**Fig. S2.5**), and this was true even when considering unique counts for pairs of genes (**Fig. S2.6**). Through extensive simulations, we demonstrated that **memento** is able to produce accurate confidence intervals for the moment estimates and well-calibrated p-values testing for their differences across groups of cells. Because our hypothesis testing framework utilizes approximate bootstrapping, it should in theory be compatible with existing parametric models and other types of estimators to enable better estimates of empirical p-values for a variety of single-cell sequencing analysis methods. For example, one could design an estimator for experiments where the mRNA sampling process cannot be approximated as a single step, and requires a more in-depth treatment. In addition, we also demonstrated that the principles behind **memento** can be extended to perform differential expression combining multiple datasets in an extremely efficient manner by frontloading expensive calculations, giving researchers better tools to interact with massive resources such as CELLxGENE Discover.

Through the application across four proof-of-principle settings, we demonstrate **memento** having increased power to detect differentially expressed genes across a range of studies. We show that our mean estimator is particularly more accurate at lower cell counts, and our inference is more concordant with results from bulk RNA-seq experiments. Moreover, we demonstrate that differential variability and correlation analysis can identify novel gene regulatory relationships that are not detected using differential mean analysis. In human tracheal epithelial cells, **memento** identified unexpected correlation of canonical ISGs at baseline, hinting at an extracellular gradient of tonic interferon and expanding the interferon response transcription regulatory network post-extracellular stimulation to encompass non-canonical ISGs. In a CD4<sup>+</sup> T cell dataset perturbed genetically by CRISPR-Cas9, **memento** analyses using genetic perturbations as causal anchors revealed genetic interactions of regulators in controlling the expression of target genes. When applied to a

population-scale scRNA-seq dataset, **memento** improved the statistical power and resolution for mapping *cis*-eQTLs, mapping additional loci impacting gene expression variability and gene correlation. Finally, we demonstrated the compatibility of the **memento** framework with the CELLxGENE Discover to power arbitrary differential expression analysis between groups of cells. Demonstrated across diverse datasets, **memento** emerges as a highly adaptable and scalable method for the quantitative analyses of large scRNA-seq datasets containing many replicates and experimental conditions.

## References

- [1] H H McAdams and A Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 94(3):814–819, February 1997. 1
- [2] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, October 2008. 1
- [3] Jitao Guo and Xuyu Zhou. Regulatory T cells turn pathogenic, September 2015. 1
- [4] John R S Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of *s. cerevisiae* reveals the architecture of biological noise. *Nature*, 441(7095):840–846, June 2006. 1
- [5] Arjun Raj, Scott A Rifkin, Erik Andersen, and Alexander Van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918, February 2010. 1, 2
- [6] Avigdor Eldar and Michael B Elowitz. Functional roles for noise in genetic circuits, September 2010. 1
- [7] Maike M K Hansen, Ravi V Desai, Michael L Simpson, and Leor S Weinberger. Cytoplasmic amplification of transcriptional noise generates substantial Cell-to-Cell variability. *Cell Syst*, 7(4):384–397.e6, October 2018. 1

- [8] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, December 2005. 2
- [9] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, April 2012. 2
- [10] Anika Gupta, Jorge D Martin-Rufino, Thouis R Jones, Vidya Subramanian, Xiaojie Qiu, Emanuelle I Grody, Alex Bloemendal, Chen Weng, Sheng-Yong Niu, Kyung Hoi Min, Arnav Mehta, Kaite Zhang, Layla Siraj, Aziz Al’ Khafaji, Vijay G Sankaran, Soumya Raychaudhuri, Brian Cleary, Sharon Grossman, and Eric S Lander. Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. *Proc. Natl. Acad. Sci. U. S. A.*, 119(34):e2207392119, August 2022. 2
- [11] Ker-Chau Li. Genome-wide coexpression dynamics: theory and application. *Proc. Natl. Acad. Sci. U. S. A.*, 99(26):16875–16880, December 2002. 2
- [12] 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, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, January 2020. 2, 72
- [13] Paul Datlinger, André F Rendeiro, Thorina Boenke, Thomas Krausgruber, Daniele Barreca, and Christoph Bock. Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing. *bioRxiv*, pages 1–27, December 2019. 2, 3, 60
- [14] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, December 2016. 2, 72

- [15] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, 36(1):89–94, January 2018. 2, 3, 26
- [16] Monique G P Van Der Wijst, Harm Brugge, Dylan H De Vries, Patrick Deelen, Morris A Swertz, and Lude Franke. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.*, 50(4):493–497, April 2018. 72
- [17] Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, Mike Thompson, Nadav Rappoport, Andrew Dahl, Cristina M Lanata, Mehrdad Matloubian, Lenka Maliskova, Serena S Kwek, Tony Li, Michal Slyper, Julia Waldman, Danielle Dionne, Orit Rozenblatt-Rosen, Lawrence Fong, Maria Dall’Era, Brunilda Balliu, Aviv Regev, Jinoos Yazdany, Lindsey A Criswell, Noah Zaitlen, and Chun Jimmie Ye. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, April 2022. 10, 13, 48, 63
- [18] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R P Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L Cook, Chun Jimmie Ye, Kirsten A Fairfax, Alex W Hewitt, and Joseph E Powell. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, April 2022. 2, 48, 63, 72
- [19] Jordan W Squair, Matthieu Gautier, Claudia Kathe, Mark A Anderson, Nicholas D James, Thomas H Hutson, Rémi Hudelle, Taha Qaiser, Kaya J E Matson, Quentin Barraud, Ariel J Levine, Gioele La Manno, Michael A Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nat. Commun.*, 12(1):

5692, September 2021. 2, 3, 10, 11, 13

- [20] Charlotte Sonesson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, 15(4):255–261, April 2018. 3
- [21] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M Keizer, Indu Khatri, Szymon M Kielbasa, Jan O Korbel, Alexey M Kozlov, Tzu-Hao Kuo, Boudewijn P F Lelieveldt, Ion I Mandoiu, John C Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J Theis, Huan Yang, Alex Zelikovsky, Alice C McHardy, Benjamin J Raphael, Sohrab P Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biol.*, 21(1):31, February 2020. 3
- [22] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.*, 53(6):770–777, June 2021. 3, 73
- [23] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, December 2018.
- [24] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, 9(1):1–17, December 2018.

- [25] Keegan D Korthauer, Li Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendzierski. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, 17(1):222, October 2016.
- [26] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, 16(1):278, December 2015.
- [27] Nils Eling, Arianne C Richard, Sylvia Richardson, John C Marioni, and Catalina A Vallejos. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell systems*, 7(3):284–294.e12, September 2018. 3, 7, 22, 27
- [28] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, Zena Werb, Eric D Chow, and Zev J Gartner. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods*, 16(7):619–626, July 2019. 3
- [29] Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Z Yeung, William M Mauck, Peter Smibert, and Rahul Satija. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.*, 19(1):224, December 2018.
- [30] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell



- transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, August 2017. 3, 73
- [31] Jingshu Wang, Mo Huang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, John Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 115(28):E6437–E6446, July 2018. 3
- [32] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018. 4, 58
- [33] Yu Fu, Pei Hsuan Wu, Timothy Beane, Phillip D Zamore, and Zhiping Weng. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, 19(1):531, July 2018. 5
- [34] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015. 7, 73
- [35] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat. Methods*, 11(6):637–640, April 2014. 7
- [36] Shila Ghazanfar, Yingxin Lin, Xianbin Su, David Ming Lin, Ellis Patrick, Ze-Guang Han, John C Marioni, and Jean Yee Hwa Yang. Investigating higher-order interactions in single-cell data with scHOT. *Nat. Methods*, 17(8):799–806, August 2020. 7
- [37] Martin Jinye Zhang, Vasilis Ntranos, and David Tse. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.*, 11(1):1–11, December 2020. 7, 15, 16, 26, 73
- [38] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert Jan Hendriks, Anton J M Larsson, Omid R Faridani, and Rickard Sandberg.

- Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.*, 38(6):708–714, June 2020. 7
- [39] Eduardo Torre, Hannah Dueck, Sydney Shaffer, Janko Gospic, Rohit Gupte, Roberto Bonasio, Junhyong Kim, John Murray, and Arjun Raj. Rare cell detection by Single-Cell RNA sequencing as guided by Single-Molecule RNA FISH. *Cell Syst*, 6(2):171–179.e5, February 2018. 8
- [40] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods*, 15(7):539–542, July 2018. 8
- [41] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, 1994. 10, 23, 24, 25
- [42] Gaia Andreoletti, Cristina M Lanata, Laura Trupin, Ishan Paranjpe, Tia S Jain, Joanne Nititham, Kimberly E Taylor, Alexis J Combes, Lenka Maliskova, Chun Jimmie Ye, Patricia Katz, Maria Dall’Era, Jinoos Yazdany, Lindsey A Criswell, and Marina Sirota. Transcriptomic analysis of immune cells in a multi-ethnic cohort of systemic lupus erythematosus patients identifies ethnicity- and disease-specific expression signatures. *Commun Biol*, 4(1):488, April 2021. 11
- [43] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J

- Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8(1):1–12, January 2017. 14
- [44] I J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953. 21
- [45] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25, March 2010. 22
- [46] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, December 2014.
- [47] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, 32(9):896–902, September 2014. 22
- [48] Aaron T L Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17(1):75, April 2016. 22
- [49] Rishi R Goel, Sergei V Kotenko, and Mariana J Kaplan. Interferon lambda in inflammation and autoimmune rheumatic diseases. *Nat. Rev. Rheumatol.*, 17(6):349–362, June 2021. 41
- [50] Liqun Zhang, Alexander Bukreyev, Catherine I Thompson, Brandy Watson, Mark E Peeples, Peter L Collins, and Raymond J Pickles. Infection of ciliated cells by human parainfluenza virus type 3 in an in vitro model of human airway epithelium. *J. Virol.*, 79(2):1113–1124, January 2005. 42
- [51] Nai-Huei Wu, Wei Yang, Andreas Beineke, Ronald Dijkman, Mikhail Matrosovich, Wolfgang Baumgärtner, Volker Thiel, Peter Valentin-Weigand, Fandan Meng, and Georg Herrler. The differentiated airway epithelium infected by influenza viruses

- maintains the barrier function despite a dramatic loss of ciliated cells. *Sci. Rep.*, 6: 39668, December 2016.
- [52] 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, Jennifer S Chen, Jin Wei, Renata B Filler, Bao Wang, Guilin Wang, Laura E Niklason, Ruth R Montgomery, Stephanie C Eisenbarth, Sidi Chen, Adam Williams, Akiko Iwasaki, Tamas L Horvath, Ellen F Foxman, Richard W Pierce, Anna Marie Pyle, David van Dijk, and Craig B Wilen. 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 Biol.*, 19(3):e3001143, March 2021. 42, 58
- [53] Daniel J Gough, Nicole L Messina, Christopher J P Clarke, Ricky W Johnstone, and David E Levy. Constitutive Type I Interferon Modulates Homeostatic Balance through Tonic Signaling, February 2012. 43, 62
- [54] Konrad C Bradley, Katja Finsterbusch, Daniel Schnepf, Stefania Crotta, Miriam Llorian, Sophia Davidson, Serge Y Fuchs, Peter Staeheli, and Andreas Wack. Microbiota-Driven Tonic Interferon Signals in Lung Stromal Cells Protect from Influenza Virus Infection. *Cell Rep.*, 28(1):245–256.e4, July 2019. 43
- [55] Tzachi Hagai, Xi Chen, Ricardo J Miragaia, Raghd Rostom, Tomás Gomes, Natalia Kunowska, Johan Henriksson, Jong-Eun Park, Valentina Proserpio, Giacomo Donati, Lara Bossini-Castillo, Felipe A Vieira Braga, Guy Naamati, James Fletcher, Emily Stephenson, Peter Vegh, Gosia Trynka, Ivanela Kondova, Mike Dennis, Muzlifah Haniffa, Armita Nourmohammad, Michael Lässig, and Sarah A Teichmann. Gene expression variability across cells and species shapes innate immunity. *Nature*, 563 (7730):197–202, November 2018. 43
- [56] Sara Mostafavi, Hideyuki Yoshida, Aviv Regev, Diane Mathis, Christophe Benoist, Devapregasan Moodley, Hugo Leboité, Katherine Rothamel, Towfique Raj,

- Chun Jimmie Ye, Nicolas Chevrier, Shen-Ying Zhang, Ting Feng, Mark Lee, Jean-Laurent Casanova, James D Clark, Martin Hegen, Jean-Baptiste Telliez, Nir Hacohen, Philip L De Jager, and Immunological Genome. Parsing the Interferon Transcriptional Network and Its Disease Associations In Brief Resource Parsing the Interferon Transcriptional Network and Its Disease Associations. 2016. 43
- [57] Eric Shifrut, Julia Carnevale, Victoria Tobin, Theodore L Roth, Jonathan M Woo, Christina T Bui, P Jonathan Li, Morgan E Diolaiti, Alan Ashworth, and Alexander Marson. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971.e15, December 2018. 44
- [58] Rachel E Gate, Christine S Cheng, Aviva P Aiden, Atsede Siba, Marcin Tabaka, Dmytro Lituiev, Ido Machol, M Grace Gordon, Meena Subramaniam, Muhammad Shamim, Kendrick L Hougen, Ivo Wortman, Su-Chen Huang, Neva C Durand, Ting Feng, Philip L De Jager, Howard Y Chang, Erez Lieberman Aiden, Christophe Benoist, Michael A Beer, Chun J Ye, and Aviv Regev. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.*, 50(8):1140–1150, August 2018. 45
- [59] Ian Dunham, Anshul Kundaje, Shelley F Aldred, Patrick J Collins, Carrie A Davis, Francis Doyle, Charles B Epstein, Seth Fretze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R Lajoie, Stephen G Landt, Bum Kyu Lee, Florencia Pauli, Kate R Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shores, Jeremy M Simon, Lingyun Song, Nathan D Trinklein, Robert C Altshuler, Ewan Birney, James B Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C Hardison, Robert S Harris, Javier Herrero, Michael M Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K Marinov, Angelika Merkel, Ali Mortazavi, Stephen C J Parker, Timothy E Reddy, Joel Rozowsky, Felix Schlesinger,

Robert E Thurman, Jie Wang, Lucas D Ward, Troy W Whitfield, Steven P Wilder, Weisheng Wu, Hualin S Xi, Kevin Y Yip, Jiali Zhuang, Bradley E Bernstein, Eric D Green, Chris Gunter, Michael Snyder, Michael J Pazin, Rebecca F Lowdon, Laura A L Dillon, Leslie B Adams, Caroline J Kelly, Julia Zhang, Judith R Wexler, Peter J Good, Elise A Feingold, Gregory E Crawford, Job Dekker, Laura Elnitski, Peggy J Farnham, Morgan C Giddings, Thomas R Gingeras, Roderic Guigó, Timothy J Hubbard, W James Kent, Jason D Lieb, Elliott H Margulies, Richard M Myers, John A Stamatoyannopoulos, Scott A Tenenbaum, Zhiping Weng, Kevin P White, Barbara Wold, Yanbao Yu, John Wrobel, Brian A Risk, Harsha P Gunawardena, Heather C Kuiper, Christopher W Maier, Ling Xie, Xian Chen, Tarjei S Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian A Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakraborty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J Luo, Eddie Park, Jonathan B Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E Antonarakis, Gregory J Hannon, Yijun Ruan, Piero Carninci, Cricket A Sloan, Katrina Learned, Venkat S Malladi, Matthew C Wong, Galt P Barber, Melissa S Cline, Timothy R Dreszer, Steven G Heitner, Donna Karolchik, Vanessa M Kirkup, Laurence R Meyer, Jeffrey C

Long, Morgan Maddren, Brian J Raney, Linda L Grasfeder, Paul G Giresi, Anna Battenhouse, Nathan C Sheffield, Kimberly A Showers, Darin London, Akshay A Bhinge, Christopher Shestak, Matthew R Schaner, Seul Ki Kim, Zhuzhu Z Zhang, Piotr A Mieczkowski, Joanna O Mieczkowska, Zheng Liu, Ryan M McDaniell, Yunyun Ni, Naim U Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R Iyer, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E Christopher Partridge, Katherine E Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M Bowling, Michael Anaya, Marie K Cross, Michael A Muratet, Kimberly M Newberry, Kenneth McCue, Amy S Nesmith, Katherine I Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L Parker, Sreeram Balasubramanian, Nicholas S Davis, Sarah K Meadows, Tracy Eggleston, J Scott Newberry, Shawn E Levy, Devin M Absher, Wing H Wong, Matthew J Blow, Axel Visel, Len A Pennachio, Hanna M Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L Tress, Marijke J Van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addleman, Roger P Alexander, Raymond K Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P Boyle, Alina R Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X Jin, Konrad J Karczewski, Maya

Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Ximeng J Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M Weissman, Luiz O Penalva, Subhradip Karmakar, Raj R Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L Bates, Rachel Byron, Theresa K Canfield, Morgan J Diegel, Douglas Dunn, Abigail K Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K Johnson, Ericka M Johnson, Tattyana V Kutuyavin, Kristen Lee, Dimitra Lotakis, Matthew T Maurano, Shane J Neph, Fiedencio V Neri, Eric D Nguyen, Hongzhu Qu, Alex P Reynolds, Vaughn Roach, Eric Rynes, Minerva E Sanchez, Richard S Sandstrom, Anthony O Shafer, Andrew B Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A Weaver, Yongqi Yan, Miaohua Zhang, Joshua M Akey, Michael Bender, Michael O Dorschner, Mark Groudine, Michael J MacCoss, Patrick Navas, George Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M Luscombe, Daniel Sobral, Juan M Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W Libbrecht, Marc A Schaub, Webb Miller, Peter J Bickel, Balazs Banfai, Nathan P Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey A Bilmes, Orion J Buske, Avinash D Sahu, Peter V Kharchenko, Peter J Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):



57–74, September 2012. 46

- [60] Aparna Nathan, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I Beynor, Yuriy Baglaenko, Sara Suliman, Alkes L Price, Leonid Lecca, Megan B Murray, D Branch Moody, and Soumya Raychaudhuri. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, 606(7912):120–128, June 2022. 47
- [61] CZI Single-Cell Biology Program, Shibli Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robotmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELL×GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. November 2023. 50
- [62] Andrew J Hill, José L McFaline-Figueroa, Lea M Starita, Molly J Gasperini, Kenneth A Matreyek, Jonathan Packer, Dana Jackson, Jay Shendure, and Cole Trapnell. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods*, 15(4):271–274, April 2018. 61
- [63] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L Bonnar, Marco Jost, Thomas M Norman, and Jonathan S Weissman. Mapping

- information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, July 2022. 72
- [64] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.*, 20(1):295, December 2019. 73
- [65] What is fixed RNA profiling? - official 10x genomics support.  
<https://www.10xgenomics.com/support/software/cell-ranger/latest/getting-started/cr-flex-what-is-frp>. Accessed: 2024-1-24. 73
- [66] Luyi Tian, Fei Chen, and Evan Z Macosko. The expanding vistas of spatial transcriptomics. *Nat. Biotechnol.*, October 2022. 73

## Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

*Mincheol Kim*

1810B016FFA9481...

Author Signature

3/16/2024

Date