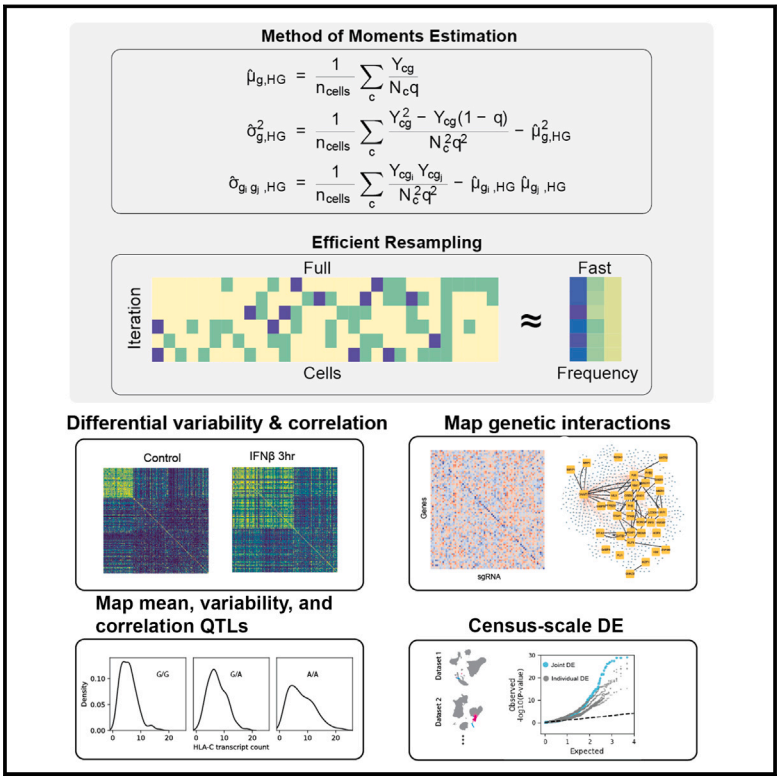
Cell

Method of moments framework for differential expression analysis of single-cell RNA sequencing data

Graphical abstract



Highlights

- A statistical model for scRNA-seq decouples measurement and expression noise
- Highly efficient resampling allows for well-calibrated hypothesis testing
- Memento enables studying coordinated expression of genes in response to perturbations
- Memento maps loci associated with gene expression mean, variability, and correlation

Authors

Min Cheol Kim, Rachel Gate, David S. Lee, ..., Alexander Marson, Vasilis Ntranos, Chun Jimmie Ye

Correspondence

jimmie.ye@ucsf.edu

In brief

Memento implements a statistical model and a fast resampling procedure to estimate and compare the mean, variability, and correlation of gene expression, allowing for the study of transcription in a deeper yet accurate fashion compared with traditional differential expression.

Kim et al., 2024, Cell *187*, 6393–6410
October 31, 2024
https://doi.org/10.1016/j.cell.2024.09.044

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Method of moments framework for differential expression analysis of single-cell RNA sequencing data

Min Cheol Kim,[1,2,3] Rachel Gate,[3] David S. Lee,[3] Andrew Tolopko,[4] Andrew Lu,[3] Erin Gordon,[5] Eric Shifrut,[6,7] Pablo E. Garcia-Nieto,[4] Alexander Marson,[7,12,13] Vasilis Ntranos,[6,8] and Chun Jimmie Ye[3,8,9,10,11,12,13,14,15,*]

[1]Medical Scientist Training Program, University of California, San Francisco, San Francisco, CA, USA
[2]UC Berkeley-UCSF Graduate Program in Bioengineering, San Francisco, CA, USA
[3]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA
[4]Chan Zuckerberg Initiative, Redwood City, CA, USA
[5]Division of Pulmonary and Critical Care, University of California, San Francisco, San Francisco, CA, USA
[6]Diabetes Center, University of California, San Francisco, San Francisco, CA, USA
[7]Division of Infectious Diseases, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA
[8]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
[9]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA
[10]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA
[11]Chan Zuckerberg Biohub, San Francisco, CA, USA
[12]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA
[13]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA
[14]Division of Rheumatology, Department of Medicine, University of California, San Francisco, San Francisco, CA, USA
[15]Lead contact
*Correspondence: jimmie.ye@ucsf.edu
https://doi.org/10.1016/j.cell.2024.09.044

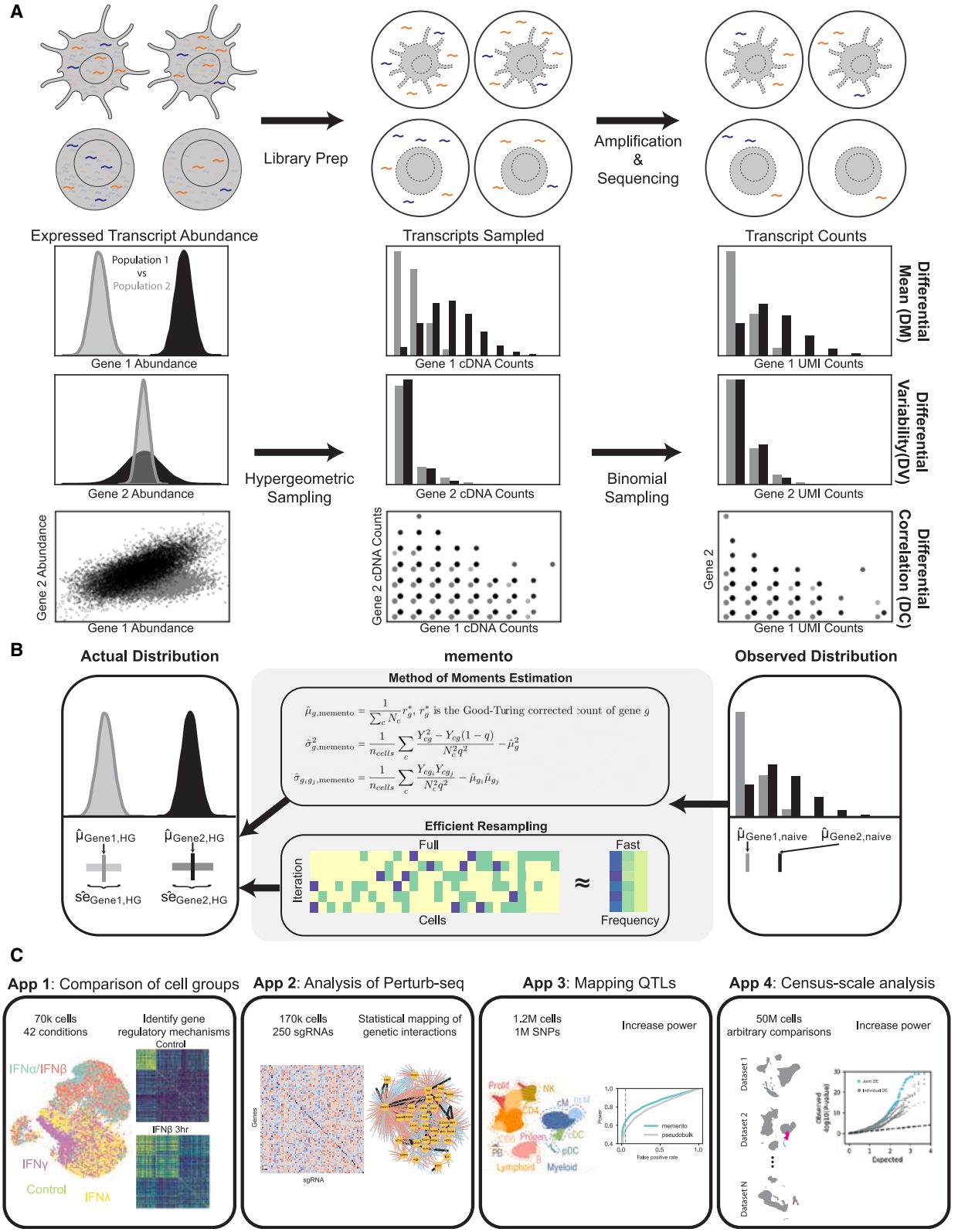
SUMMARY

Differential expression analysis of single-cell RNA sequencing (scRNA-seq) data is central for characterizing how experimental factors affect the distribution of gene expression. However, distinguishing between biological and technical sources of cell-cell variability and assessing the statistical significance of quantitative comparisons between cell groups remain challenging. We introduce Memento, a tool for robust and efficient differential analysis of mean expression, variability, and gene correlation from scRNA-seq data, scalable to millions of cells and thousands of samples. We applied Memento to 70,000 tracheal epithelial cells to identify interferon-responsive genes, 160,000 CRISPR-Cas9 perturbed T cells to reconstruct gene-regulatory networks, 1.2 million peripheral blood mononuclear cells (PBMCs) to map cell-type-specific quantitative trait loci (QTLs), and the 50-million-cell CELLxGENE Discover corpus to compare arbitrary cell groups. In all cases, Memento identified more significant and reproducible differences in mean expression compared with existing methods. It also identified differences in variability and gene correlation that suggest distinct transcriptional regulation mechanisms imparted by perturbations.

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.[1,2] While genetics and environmental history significantly contribute to expression variability across a population of cells, stochastic transcriptional noise can also influence cellular responses to perturbations, as well as cellular development and differentiation.[2–4] Characterizing how deterministic and stochastic factors jointly influence the distribution of gene expression is central to understanding how transcriptional control is established, maintained, and may be broken. These insights could illuminate mechanisms underlying phenomena where genotype-phenotype relationships are not completely explained, such as destabilization,[3] incomplete penetrance,[5] and variable expressivity.[6]

The distribution of gene expression within a population of cells is primarily characterized by its mean and variance and related derived measures.[7] Constitutively expressed housekeeping genes, which undergo transcription and degradation at constant rates, are predicted to conform to a Poisson distribution. Nonetheless, most genes display over-dispersion, exhibiting higher variance than expected,[8] and genes within the same biological pathway are often transcriptionally correlated.[5] These



(legend on next page)

observations are consistent with a model where the expression of related genes is regulated by similar *cis*-regulatory elements that interact with a common set of transcription factors that cycle between "on" and "off" states.[9] 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.[10,11]

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,[12] genetic perturbations,[13,14] and natural genetic variation.[15–18] In theory, the analysis of scRNA-seq data can reveal how deterministic and stochastic factors together shape the distribution of gene expression. Yet, there remains a need for differential 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 perform differential expression analysis on pseudobulk profiles, generated by aggregating transcript counts for cell groups defined by clustering. While pseudobulk approaches do not fully leverage single cells as repeated measures, they surprisingly outperform methods that explicitly model the distribution of observed scRNA-seq data.[19] Moreover, very few methods exist for assessing differences in gene expression variability and correlation between pairs of genes.

Generalized differential expression analysis of scRNA-seq data remains a formidable challenge due to two pivotal statistical limitations. First, decomposing the observed cell-to-cell variability into its constituent components—biological and measurement noise—presents a significant obstacle.[20] This difficulty stems from the small numbers of molecules involved in the biochemical reactions of both gene transcription and the scRNA-seq sampling process (Figure 1A).[21] Most existing methods implement parameterized models designed to account for the higher-than-expected variance in the *observed* sparse transcript counts. However, these models do not explicitly model measurement noise, a byproduct of the inherent undersampling characteristics of scRNA-seq workflows.[22–27] Importantly, accurately estimating biological variability is crucial for effectively modeling the correlation between gene pairs.[22] Second, establishing the statistical significance of a specific comparison of mean, variability, or gene correlation between cell groups remains a largely unsolved problem. Many existing methods utilize asymptotic theory to determine the significance of hypothesis tests comparing means, often producing uncalibrated p values. This is particularly problematic for studies necessitating thousands of comparisons, as inadequately calibrated p values violate assumptions for multiple testing correction. Moreover, most existing methods require an exact specification of the parametric model and lack flexibility in incorporating hierarchical structures and continuous covariates effectively. Thus, they do not explicitly account for biological and technical replicates inherently generated from multiplexed workflows that accommodate a growing number of individuals or conditions.[13,15,28–30] Methods like DESCEND, which utilize flexibly defined generalized linear models, are notable exceptions and are theoretically equipped to effectively address this issue. However, these models often encounter significant computational hurdles when modeling the complex hierarchical structure inherent in scRNA-seq data and are limited to a specific model of cell-cell variability.[31] Indeed, recent studies have reported a startling underperformance of scRNA-seq methods relative to pseudobulk methods when testing mean differences.[19]

To address 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 (Figure 1B). Memento employs a multivariate hypergeometric sampling process and leverages the sparsity of scRNA-seq data to implement a bootstrapping strategy for the efficient statistical comparisons 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 and genetic factors affect the distribution of gene expression in human cells (Figure 1C). First, we conducted scRNA-seq on 70,000 tracheal epithelial cells stimulated with extracellular interferons (IFNs) and investigated how stimulation modulates the variability and correlation of response genes temporally. Second, we performed Perturb-seq on 170,000 T cells and mapped gene-regulatory networks that define aspects of broad T cell activation. Third, we reanalyzed 1.2 million 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

Figure 1. Memento workflow for differential mean, variability, and gene correlation testing

(A) Experimental workflow for single-cell RNA sequencing (scRNA-seq) samples RNA transcripts inside each cell during library preparation and sequencing. After scRNA-seq sampling, patterns of mean, variability, and correlation of gene expression in the observed transcript counts no longer resemble the actual distribution.

(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 efficient bootstrapping for estimating confidence intervals (CIs), and tests for differences in expression parameters between two groups of cells.

(C) Four applications of Memento to characterize the response of ~70,000 human tracheal epithelial cells to extracellular cytokines, reconstruct gene-regulatory networks from ~170,000 human CD4+ T cells perturbed by CRISPR-Cas9, map the genetic determinants of gene expression in 1.2 M peripheral blood mononuclear cells (PBMCs) from 162 systemic lupus erythematosus (SLE) patients and 99 healthy controls, and comparisons of arbitrary groups of cells within the CELLxGENE Discover data corpus.

See also Figure S1.

Discover Census Application Programming Interface (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 with existing methods. It also identified differences in expression variability and gene correlation, thereby revealing distinct modes of transcriptional regulation imparted by perturbations. Memento is implemented in Python, is compatible with scanpy,[32] and can be downloaded at https://github.com/yelabucsf/scrna-parameter-estimation.

RESULTS

Statistical model of scRNA-seq

Since its advent, 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 (Figure 1A). Decomposing this variability into components of biological and measurement noise is pivotal for differential expression analysis of scRNA-seq data.

Here, we propose a 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 reverse transcription (RT) and sequencing of resultant libraries are processes that 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 $\boldsymbol{X}_c = \frac{\boldsymbol{Z}_c}{N_c}$ denote an m -dimensional random variable representing the normalized transcript counts of m genes in cell c, where $\boldsymbol{Z}_c$ 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 $\boldsymbol{Y}_c$ originate from $\boldsymbol{X}_c$: $\boldsymbol{Y}_c \sim \text{MultiHG}(N_c\boldsymbol{X}_c, N_c, N_c q)$. 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 STAR Methods for detailed exploration). Importantly, we empirically substantiate that the two-step noise process involving RT (hypergeometric) and sequencing (binomial) can be well represented with a single step of hypergeometric sampling with the overall q (Figure S1A). Across many simulated values of capture efficiency and sequencing saturation, the single-step hypergeometric sampling closely approximates the two-step process (nonsignificant Kolmogorov-Smirnov test; Figure S1B).

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 $\boldsymbol{X}_c$ given $\boldsymbol{Y}_c$ under the assumption of hypergeometric sampling (see STAR Methods for derivation and details):

$$\widehat{\mu}_{g,\text{memento}} = \frac{1}{\sum_c N_c} r_g^*, \text{ where } r_g^* \text{ is the Good-Turing corrected count of gene } g$$

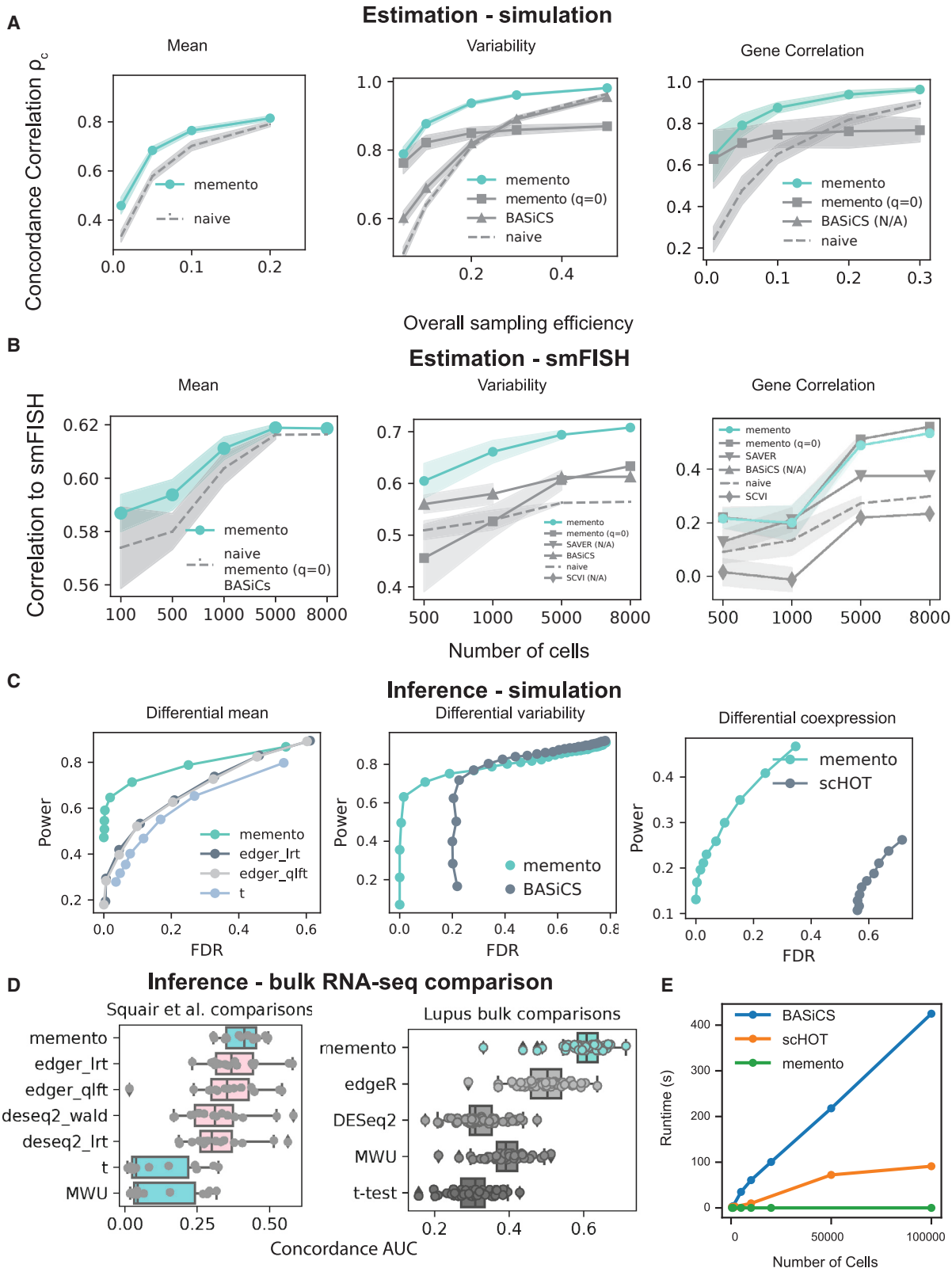
$$\widehat{\sigma}^2_{g,\text{memento}} = \frac{1}{n_{cells}} c \frac{Y_{cg}^2 - Y_{cg}(1-q)}{N_c^2 q^2} - \widehat{\mu}_g^2$$

$$\widehat{\sigma}_{g_i g_j,\text{memento}} = \frac{1}{n_{cells}} \sum_c \frac{Y_{cg_i} Y_{cg_j}}{N_c^2 q^2} - \widehat{\mu}_{g_i} \widehat{\mu}_{g_j}$$

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.[33,34] To do so, we introduce the residual variance $\tilde{\sigma}_g$ as a measure of expression variability σ_g (STAR Methods), defined as the variance component unexplained by the mean (STAR 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,[35] empirical Bayes estimators under the Poisson approximation introduced by Zhang et al.[36] (a special case of the Memento estimator for setting $q = 0$), and estimates derived from BASiCS[27] (see STAR Methods for forms of the naive and Poisson estimators). Across a range of q values, Memento's hypergeometric estimator produced accurate estimates of mean (Lin's concordance correlation coefficient—$\rho_c > 0.8$ with 10 cells and > 0.98 with 100 cells), residual variance ($\rho_c > 0.98$, 100 cells), and gene correlation ($\rho_c > 0.98$, 100 cells) (Figure 2A). In addition, Memento produces stable residual variance and gene correlation estimates across q s, outperforming other estimators for both low- and high-efficiency scRNA-seq workflows. While all estimators have higher accuracy for highly expressed genes, Memento outperforms other methods even for lowly expressed genes (Figure S1C). 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 (DropSeq) and single-molecule fluorescence *in situ* hybridization (smFISH) data.[37] This dataset was previously analyzed using SAVER,[38] an imputation method that borrows information from similar genes and cells for estimating gene correlations that has been shown to outperform other approaches (Figure 2B). For genes profiled using both DropSeq and smFISH, Memento's mean estimates exhibited modest improvements over the naive estimator used by other methods when very few cells are used (21 genes considered; $\rho = 0.58$ and $\rho = 0.54$, using 100 cells). For residual variance, Memento's estimates were



(legend on next page)

significantly more correlated with those obtained by smFISH (14 genes considered; ρ = 0.71) than the naive estimator (ρ = 0.56) and BASiCS (ρ = 0.61) using all available 8,498 cells. Finally, for gene correlation, Memento (ρ = 0.53) also significantly outperforms the naive estimator (ρ = 0.29), SAVER (ρ = 0.38), and scVI (ρ = 0.23) 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 s vs. SAVER, 30 min 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. These results underscore the accuracy of Memento's parameter estimates, demonstrated through both simulations and comparative analyses against benchmark smFISH data.

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. A 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, the appropriate calibration of the test statistics under the null hypothesis amenable to multiple testing correction becomes imperative. Although employing MoMs estimation offers computational efficiency and modeling flexibility, establishing the statistical significance of estimated parameters necessitates the computation of confidence intervals (CIs) through bootstrapping of 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 a shuffling scheme that capitalizes on the sparsity of scRNA-seq data to facilitate fast, memory-efficient, and highly parallelizable bootstrapping. Our scheme is based on the key observation that the number of unique observed transcript counts is substantially smaller than the number of cells (Figure S2A), and this held true even for unique observed pairs of counts (Figure S3B), albeit to a lesser extent. Therefore, each bootstrap iteration necessitates merely the resampling of K unique transcript counts for each gene from Multinomial$\left(N, \frac{n_1}{N} \ldots \frac{n_K}{N}\right)$, proportional to the observed frequency of each count (Figure S2C), as opposed to resampling individual cells' counts from a multinomial distribution comprising N elements (cells) (Multinomial$\left(N, \frac{1}{N} \ldots \frac{1}{N}\right)$[39]). This approach culminates in fitting a markedly small weighted dataset ($K \ll N$) for each resampling iteration. To accommodate multiplexed experiments, we extend our boostrapping 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 were 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 (Figures S2D and S2E). Utilizing bootstrapping to quantify the CI in parameter estimates, Memento computes well-calibrated empirical p values for DM, DV, and differential correlation (DC), suitable for multiple testing correction (Figure S2F).

To show that Memento produces accurate estimates of false positives 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 CD4 + T cells pre- and post-stimulation with recombinant IFN-β (rIFNB) were employed. We show that for DM, DV, and DC, Memento estimated the expected number of false positives at a specified significance cutoff while achieving the highest power for detecting true parameter differences (Figure 2C). Moreover, we observed that existing scRNA-seq DM methods are too liberal (t test, Wilcoxon rank-sum test) while pseudobulk DM methods are far too conservative (edgeR, DESeq2), consistent with results from Squair et al.[19] (Figure S3A). Squair et al. previously attributed this result to replicate-level heterogeneity present in most scRNA-seq datasets and recommended pseudobulk methods to simplify the hierarchical structure.[19] By directly accounting for the hierarchical structure, Memento produced expected false positive rates (FPRs) at each significance threshold even when varying degrees of heterogeneous 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.[19] (Figure 2D, left) and an additional dataset from systemic lupus erythematosus (SLE) patients (Figure 2D, right).[40] In both datasets, Memento produced DM results from the scRNA-seq data most concordant with those obtained from analyses of bulk RNA-seq. Finally, Memento identified the greatest number of concordantly

Figure 2. Performance of Memento in simulation and on real data

(A) Lin's concordance of estimates of mean using 10 cells (left), variability using 100 cells (middle), and gene correlation using 100 cells (right) with simulated ground truth values (y axis) for a range of overall transcript capture efficiencies (x axis). Shaded region indicates the standard error.

(B) Pearson correlation of Memento estimates from DropSeq data vs. smFISH estimates of the same population of melanoma cells (y axis) for mean (left), variability (middle), and gene correlation (right) across different numbers of DropSeq cells used (x axis). Shaded region indicates the standard error.

(C) Power (y axis) vs. false discovery rate (FDR) (x axis) comparing existing methods with Memento for DM (left), DV (middle), and DC (right) analyses.

(D) Concordance Area Under the Curve (AUC) (x axis) of single-cell DM analysis (green) compared with pseuobulk DM analysis (red) using datasets in Squair et al.[19] and Perez et al.[17]

(E) Runtime (y axis) of three methods across number of cells (x axis) for DM and DV analyses.

See also Figures S2 and S3.

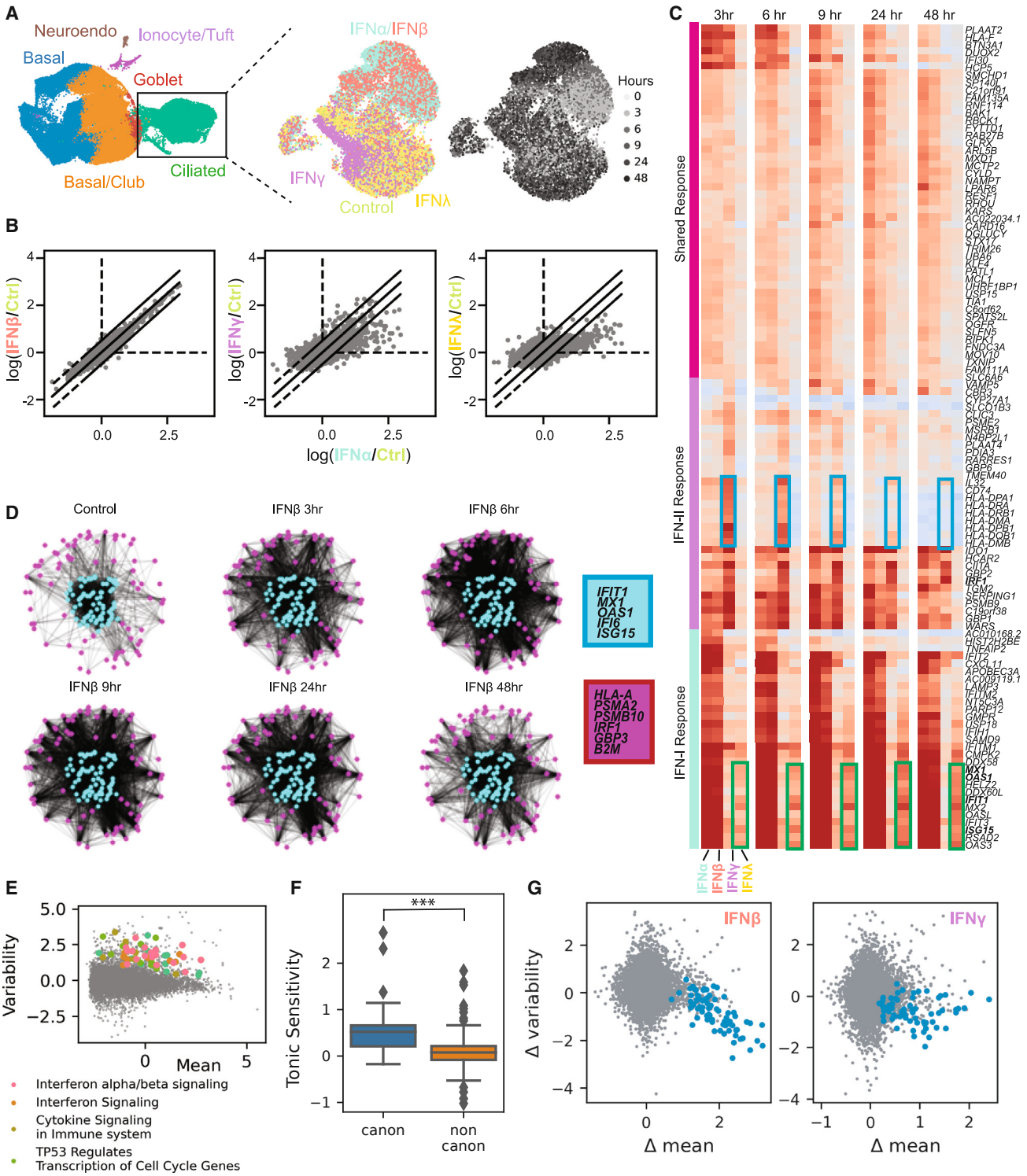


Figure 3. Mapping transcriptional response of HTECs to extracellular interferon using Memento

(A) UMAPs of the entire HTEC 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-α (x axis) against LFC in response to IFN-β (left), IFN-γ (middle), and IFN-λ (right) after 6 h.

(C) Hierarchically clustered heatmaps of LFC in response to the four types of interferons (columns within each heatmap) across 5 time points. Type-1- (green) and type-2-specific (blue) responses are highlighted.

(legend continued on next page)

differentially expressed genes in ciliated cells stimulated by IFN-α and IFN-β, both of which are known to be ligands of the type-1 IFN receptor (Figure S3B).

Compared with existing methods for DM, DV, and DC, Memento achieves hypothesis testing at computational speeds orders of magnitude faster, allowing scalability to millions of cells (Figure 2E). In a simulation comparable in scale to emerging scRNA-seq datasets (two groups each containing 10^6 cells) conducting DM and DV analyses for 1,000 genes using 10,000 bootstrapping iterations per gene required only 13 min using a single CPU. A multicore implementation of Memento facilitated the parallelization of multiple genes, further reducing the runtime to 2–3 min with 6 CPUs. Particularly for DV and DC analyses, Memento achieves computational speed gains up to 1,000× using equivalent compute resources compared with existing methods. These results substantiate that Memento's bootstrapping strategy yields accurate CI estimates for effect sizes at high computational efficiency. This culminates in well-calibrated test statistics, facilitating hypothesis testing of scRNA-seq data scalable to groups containing millions of cells (see STAR Methods for detailed description of the resampling strategy and hypothesis testing).

Differential variability and gene correlation in response to exogenous IFN

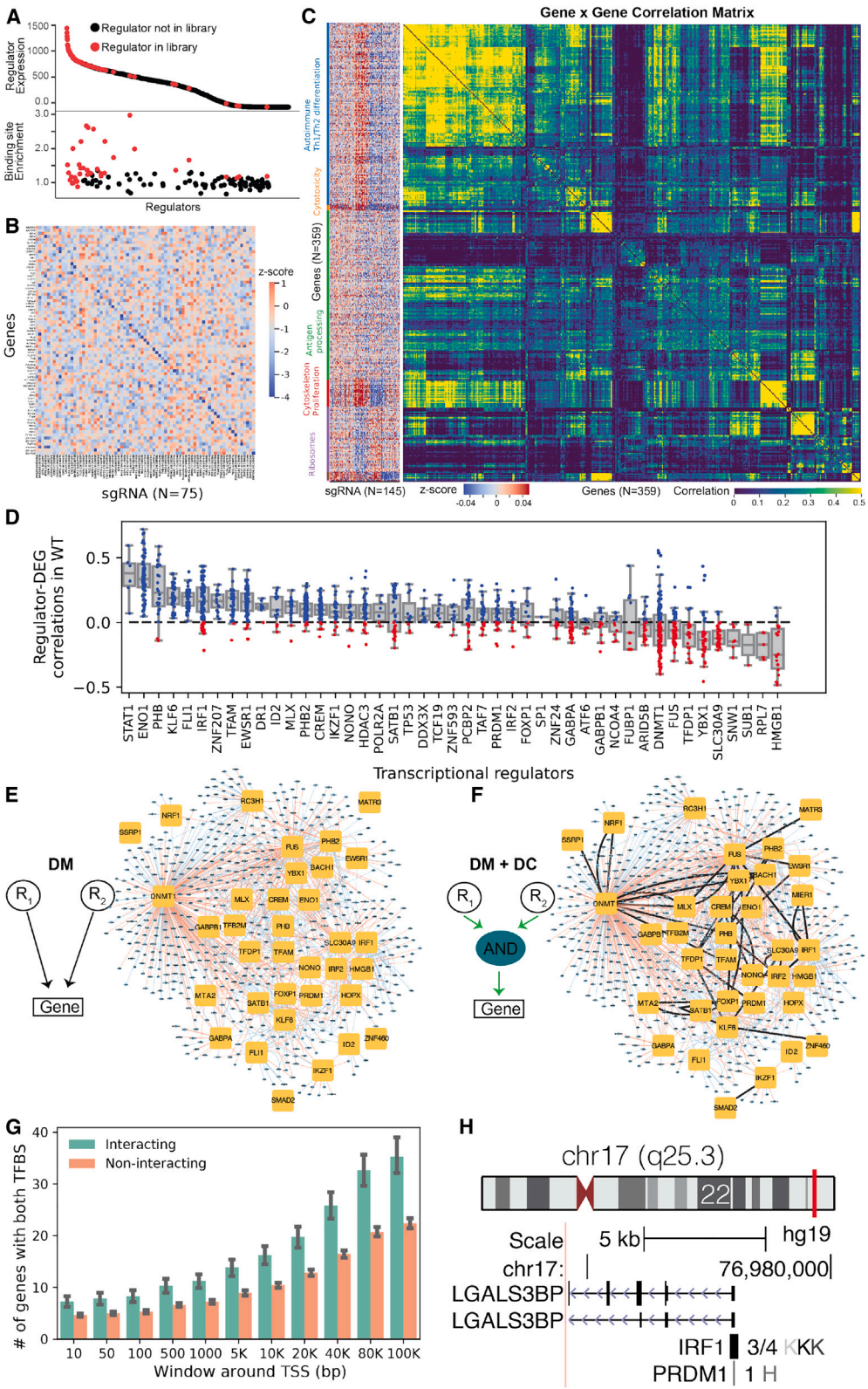
While IFNs are potent cytokines that promote antiviral immunity, they also play a role in the pathogenesis of inflammatory and autoimmune diseases.[41] Their action—inducing gene expression via autocrine and paracrine signaling—is well documented; however, the heterogeneity of transcriptomic responses in stimulated cells remains largely unexplored. Using Memento, we investigated the impact of IFN stimulation on the distribution of gene expression in human tracheal epithelial cells (HTECs). We used multiplexed scRNA-seq (mux-seq) to analyze 69,958 HTECs from two healthy donors, exploring conditions including unstimulated control and stimulation with various IFNs: type-1 (IFN-α and IFN-β), type-2 (IFN-γ), and type-3 (IFN-λ). Analyses were conducted at several post-stimulation time points: 3, 6, 9, 24, and 48 h. 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 (Figure 3A). Our subsequent analyses 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 IFN response.[42–44]

We identified 5,018 genes exhibiting differential mean expression (DMGs, false discovery rate [FDR] < 0.01) between unstimulated ciliated cells and those stimulated by any of four IFNs at 6 h. A comparative analysis revealed that IFN-α induces similar fold changes (FCs) across DMGs compared with IFN-β and IFN-λ ($\rho = 0.96$). By contrast, compared with IFN-γ, the overall correlation in FC was lower ($\rho = 0.70$; Figure 3B) due to the presence of both type-1 and type-2 IFN-specific DMGs. Herein, we define DMGs that are upregulated in response to any IFN as IFN-stimulated genes (ISGs). Hierarchical clustering of the ISGs across time points revealed a dynamic transcriptomic response shared across IFNs, including the early induction of major histocompatibility complex (MHC) class II genes and a distinct gene cluster, comprising *PLAAT2*, *BTN3A1*, and *DUOX2* (Figure 3C). We also identified patterns specific to each IFN, exemplified by a subset of canonical ISGs (*IFIT2*, *IFITM2*, and *ISG15*) that exhibited late induction in response to IFN-λ but sustained induction throughout the time course in response to type-1 IFNs (Figure 3C). Interestingly, some genes that were more induced by one of the IFNs (e.g., the MHC class II genes by IFN-γ) showed similar temporal behavior across the other IFNs, suggesting both unique and shared regulatory mechanisms.

While DM analysis 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 map the IFN gene correlation network and its subcomponents, we used Memento to identify DC between ISG pairs across stimulations and time points (Figure 3D). Agglomerative clustering of the resulting gene correlation matrix revealed distinct ISG subsets in response to IFN-β, forming clusters in unstimulated cells, stimulated cells, or both—distinctions that were not discernible through DM analysis alone. For example, canonical ISGs, including *MX1*, *OAS1*, and *IFI6*, maintained high correlation even without exogenous IFN presence (Figure 3D, cyan nodes). Upon IFN-β 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 (Figure 3D, magenta nodes). Consistent with the clustering analysis, Memento identified more differentially correlated gene pairs (DCGs, FDR < 0.1) among non-canonical ISGs (860 DCGs, 34% of total pairs) than canonical ISGs (421 DCGs, 16% of total pairs). Notably, the increase in correlation between gene pairs was not explained by an increase in their mean expression when considering all pairs of genes and when only considering pairs with significant changes in mean (Figure S4A).

We hypothesized that canonical ISGs are correlated in unstimulated cells due to the sensing of tonic IFN and the coordinated induction of ISGs within a select group of cells. Tonic IFN has been described as inducing a natural gradient of ISG expression across cells[45,46] and plays an important role in viral defense,[46] immune cell homeostasis, and autoimmunity.[45] Within our dataset, canonical ISGs exhibited greater variability compared with non-canonical ISGs in unstimulated cells (Figure 3E), aligning

(D) Gene coexpression network over time, where cyan nodes depict canonical ISGs and magenta nodes depict non-canonical ISGs. Pairs of genes with high correlation (Memento $\rho > 0.6$) are connected.
(E) Baseline expression variability (y axis) vs. mean (x axis) in ciliated cells.
(F) Tonic sensitivity (y axis) for canonical and non-canonical ISGs (x axis). ***$p < 0.001$.
(G) Change in variability (y axis) vs. the change in the mean (x axis) in response to IFN-β (left) and IFN-γ (right). Blue dots represent canonical ISGs.
See also Figure S4.



(legend on next page)

with previously documented differences in expression variability between cytokines and non-cytokines (Figure S4B).[47] 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 IFNs at 6 h, 394 were highly variable in unstimulated cells (FDR < 0.005) and were enriched for canonical ISGs (GSEA IFN-α/IFN-β signaling adjusted $p = 3.35 \times 10^{-12}$), including *IFIT1*, *IFIT3*, and *MX1*.

We next assessed the sensitivity of each ISG to tonic IFN, estimated as the FC in gene expression between macrophages from *Ifnar* knockout and wild-type (WT) mice without exogenous IFN.[48] This analysis revealed that canonical ISGs are significantly more sensitive to tonic IFN than non-canonical ISGs ($p\ P < 2.73 \times 10^{-10}$; Figure 3F). Notably, upon stimulation with IFN-β (and, to a lesser extent, with IFN-γ), the variability of substantial proportion of canonical ISGs reduced (78% and 39%, respectively; Figure 3G, FDR < 0.1), implying that exogenous stimulation homogenizes the cellular environment, removing the effects of heterogeneous response to tonic IFN.

Our findings underscore the power of Memento to analyze gene expression distributions and uncover transcriptional regulatory networks influenced by IFN signaling. By leveraging Memento to dissect effects on mean, variance, and correlation in gene expression, we have illuminated complex regulatory interactions that dictate cellular behavior in the presence and absence of IFN, offering new perspectives on how cells modulate their transcriptomic response to environmental cues.

Differential expression analysis of perturbed CD4$^+$ 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 genetic screens in diverse *in vitro* systems. Utilizing Memento, we analyzed ~173,000 CRISPR-Cas9 perturbed human CD4$^+$ T cells to map transcriptional regulatory networks modulating their activation and polarization. Cells were perturbed using pooled single-guide RNA (sgRNA) lentiviral infection with Cas9 protein electroporation (SLICE),[49] followed by mux-seq.[15] 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 assay for transposase-accessible chromatin with sequencing [ATAC-seq]) in activated CD4$^+$ T cells[50] (Figure 4A). After Cas9 electroporation and multiple rounds of selection and proliferation, activated CD4$^+$ 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 of sequencing coverage of edited cells at the target locus to sequencing coverage of its respective sgRNA in the pool, was estimated at 21%, with a standard deviation of 15% (Figure S5A). Fourteen sgRNAs, exhibiting cutting efficiencies below 2.0% (standard deviation 1.7%; Z score, $p\ P < 0.05$), 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 that target genes in cells transduced with the respective sgRNA were significantly downregulated (Figure 4B). Second, a higher correlation in average gene expression was observed between either WT cells ($\rho = 0.50$) or cells transduced with sgRNAs targeting the same gene ($\rho = 0.44$), as compared with cells transduced with sgRNAs targeting two distinct genes ($\rho = 0$; KS test $P < 2.2\ x\ 10^{-16}$ for both; Figure S5B).

Utilizing Memento, we identified 7,641 genes (FDR < 0.05) with DMGs when comparing WT cells with cells perturbed by at least one sgRNA. Hierarchical clustering of mean gene expression for DMGs across sgRNAs revealed clusters of sgRNAs exerting similar transcriptomic effects and clusters of genes similarly responsive to such perturbations (Figure 4C). We identified five clusters of DMGs distinctly associated with ribosomes (FDR < 5.35×10^{-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 (Figure 4C). Intriguingly, while antigen processing genes' mean expression is modulated by a shared set of sgRNAs, a subset of MHC class II genes—namely *HLA-DPA1*, *HLA-DRA*, *HLA-DRB1*, and *HLA-DPB1*—exhibited strong correlation, suggesting that their coordinated 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 TRs might be possible without conducting combinatorial perturbations. To test this hypothesis, we performed a genetic interaction analysis focused on pairs consisting of DMGs and their TRs, referred to as TR-DMGs (see STAR Methods). Specifically, we focused on regulators that, when

Figure 4. Reconstructing gene-regulatory networks of T cell activation using Perturb-seq and Memento

(A) Selection criteria for perturbed regulators in this study, based on expression (top) and binding site enrichment (bottom).

(B) Heatmap of average gene expression for each gene (row) across cells perturbed by the corresponding sgRNA (columns).

(C) Left: heatmap of average gene expression for DMGs (row) across cells perturbed by each sgRNA (columns). Right: gene-gene correlation matrix for the same DMGs estimated from WT cells.

(D) Correlation between each regulator and its downstream targets in WT cells.

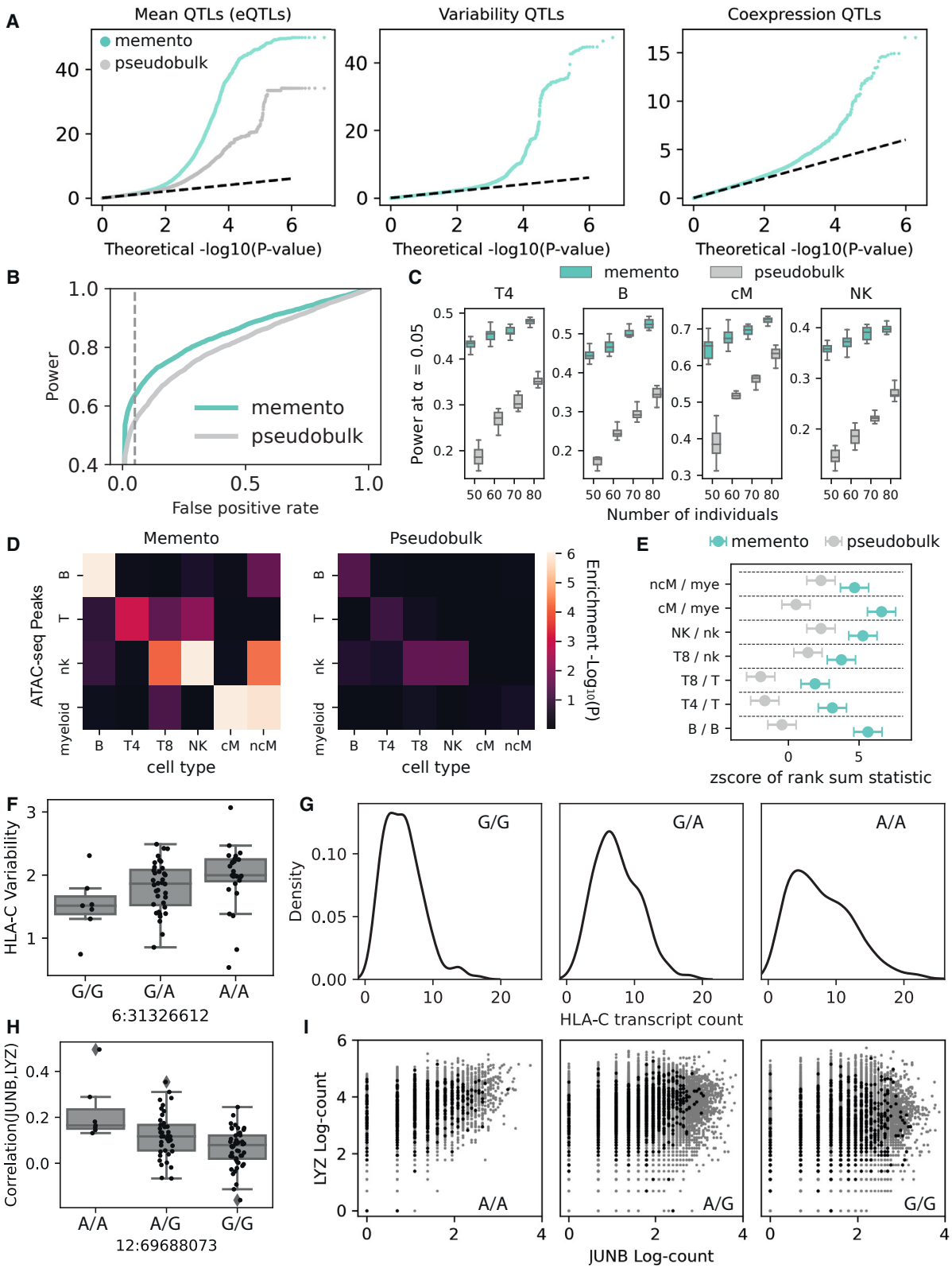
(E) Bipartite gene-regulatory network that do not account for interactions between regulators constructed from DM analysis of Perturb-seq data.

(F) Gene-regulatory network including genetic interactions between regulators constructed utilizing both DM and DC analysis.

(G) Number of genes with binding sites for pairs of interacting or non-interacting regulators across varying windows around the TSS. Error bar indicates the standard error.

(H) Chromosomal location of *LGALS3BP* and binding sites for IRF1 and PRDM1, predicted to interact using DM and DC analysis.

See also Figure S5.



(legend on next page)

knocked out, lead to decreased expression of the DMGs. Consistent with our expectations, TR-DMGs typically show a positive correlation with each other within WT cells (Binomial test, $p < 0.00668$; Figure 4D).

In the absence of a genetic interaction, two TRs (R1 and R2) could independently regulate the target gene (G); therefore, a knockout of one regulator should ostensibly not impair the functionality of the other (Figure 4E). By contrast, in the presence of an interaction, 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 (Figure 4F). Employing this strategy, we identified 564 genetic interactions amidst 432 unique regulator pairs (FDR < 0.1, Figure 4F). Validating these interactions, analyses incorporating chromatin immunoprecipitation sequencing (ChIP-seq) data from ENCODE[51] 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 (Figure 4G).

As an example, we identified that *IRF1* regulates *LGALS3PB* (evident from DM expression analysis) and retains a strong correlation with *LGALS3PB* 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 (Figure 4H).

These results demonstrate the capability of Memento for the analyses of forward genetic Perturb-seq screens. We highlight the potential for DC analyses in delineating gene sets sharing regulatory elements—albeit participating in diverse pathways—and to reconstruct the genetic interactions of *trans* regulators orchestrating T cell activation.

Genetic analysis of population-scale scRNA-seq

The increasing 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 expression quantitative trait loci (eQTLs), to identify *cis* eQTLs (*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.[52] We posit that, in comparison to pseudobulk methods, Memento's superior parameter estimation accuracy and capacity to account for intra- and inter-individual variation during inference will result in increased power to detect *cis*-eQTLs and the discovery of novel variability and correlation QTLs (vQTLs and cQTLs, 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 1.2 million peripheral blood mononuclear cells (PBMCs) derived from 162 SLE patients and 99 healthy donors.

The data were analyzed separately for each of the reported cell types: CD4 + T cells (T4), CD8 + T cells (T8), natural killer (NK) cells, classical monocytes (cMs), and non-classical monocytes (ncMs).[17] 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 100 kb from the TSS—associated with mean expression, expression variability, and gene correlation, producing well-calibrated *p* values (Figure 5A).

A comparative analysis between the power and FPR of Memento and Matrix eQTL in detecting *cis*-eQTLs was established against benchmarks provided by the OneK1K study, which comprised of 1,000 non-overlapping individuals.[18] 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 (Figures 5A and 5B). 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 mux-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 (Figure 5C).

We subsequently explored whether the increased number of *cis*-eQTLs detected by Memento also improves the enrichment

Figure 5. Mapping of mean QTL (eQTL), vQTL, and cQTL using Memento

(A) Quantile-quantile (QQ) plots for expected *p* values (y axis) computed by Memento vs. theoretical *p* values (x axis) for eQTLs, vQTLs, and cQTLs. For eQTLs, the QQ-plot of *p* values from pseudobulk approach (matrix eQTL) is overlayed.

(B) Receiver operating characteristic (ROC) curves for recovery of eQTLs identified from a much larger cohort (OneK1K) for Memento and pseudobulk-based matrix eQTL.

(C) Power of eQTL recovery (y axis) of Memento and matrix eQTL across different numbers of individuals. Analyses were performed on CD4 + T cells (T4), B cells (B), classical monocytes (cMs), and natural killer cells (NKs).

(D) Enrichment of cell-type-specific eQTLs in cell-type-specific ATAC peaks. Each entry represents the enrichment for eQTLs detected in one cell type (column) in ATAC peaks detected in another cell type (row). Intensity is $-\log_{10}(p \text{ value})$.

(E) Enrichment of eQTLs detected in each cell type for cell-type-specific ATAC peaks detected in the same cell type. Error bar indicates the standard error.

(F) An example of a vQTL. Expression variability (y axis) for each individual of varying genotypes at chr6:31326612.

(G) Histogram showing distribution of *HLA-C* expression for a representative individual of each genotype.

(H) An example of a cQTL. JUNB-LYZ gene correlation (y axis) for individuals of varying genotypes at chr12:69688073.

(I) Scatterplot of expression of *LYZ* (y axis) against the expression of *JUNB* (x axis) across single cells from all donors (gray) and a representative individual (black).

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, 9.0×10^{-9} vs. 0.04; T4, 9.3×10^{-4} vs. 0.11; T8, 0.03 vs. 0.58; NK, 6.67×10^{-8} vs. 0.03; cM, 2.1×10^{-11} vs. 0.67; ncM, 1.0×10^{-6} vs. 0.46; Figures 5D and 5E). Similar gains in enrichment were observed in the European cohort (Figure S5C). Linkage disequilibrium (LD) score regression (LDSC) analysis found that *cis*-eQTLs identified by Memento also were more enriched for genome-wide association study (GWAS) associations to immune-mediated diseases, thereby suggesting improved fine-mapping performance (Figure S5D).

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 vQTLs impacting 733 genes across all cell types. For instance, the variability in *HLA-C* expression differed among genotypes of chr6:31326612 (Figure 5F), with the A allele amplifying the expression variability of *HLA-C* without notably affecting its mean (Figure 5G). For mapping 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 chr12:69688073 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 1 kbp 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 unveil novel mechanisms whereby genetic variants may modulate gene expression.

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 diverse datasets, including the analysis of the temporal response of tracheal epithelial cells stimulated by IFNs, the mapping of gene-regulatory networks from Peturb-seq data of CD4+ T cells, and large-scale genetic analysis of scRNA-seq data collected across a population cohort. These applications and simulations demonstrate that Memento consistently outperforms existing methods, delivers a unique feature set to compare variances and correlations in addition to means, and is extremely efficient, allowing for scalability to millions of cells and tens of thousands of replicates.

The emergence of massive repositories of scRNA-seq data worldwide has created new demands 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 and thousands of individuals, with its Census API providing access to most of these data.[53] 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 type residing in different organ systems. Another user may be interested in differences in expression for the same cell type between individuals of different disease statuses. In any scRNA-seq dataset with labeled cell types, there is a large number of possible comparisons between cell groups (Figures 6A and 6B). 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 Figure 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 CZI to utilize the CELLxGENE Discover Census API to perform bootstrap operations and quantify uncertainty for predefined cell groups across the entire corpus (see STAR 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 cMs from a single donor in the lupus dataset (referenced in Figure 5), also included in 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. Our expectations were confirmed by observing a robust correlation in the effect size estimates (Figure S6) between the full and approximate, precomputed modes. A similarly strong correlation was noted in the significance levels, indicated by $-\log_{10}(p \text{ value})$ (Figures 6C and 6D). Importantly, the computation time for determining effect size and *p* value was significantly reduced compared with executing Memento in full mode for various cell group comparisons (Figure 6E).

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

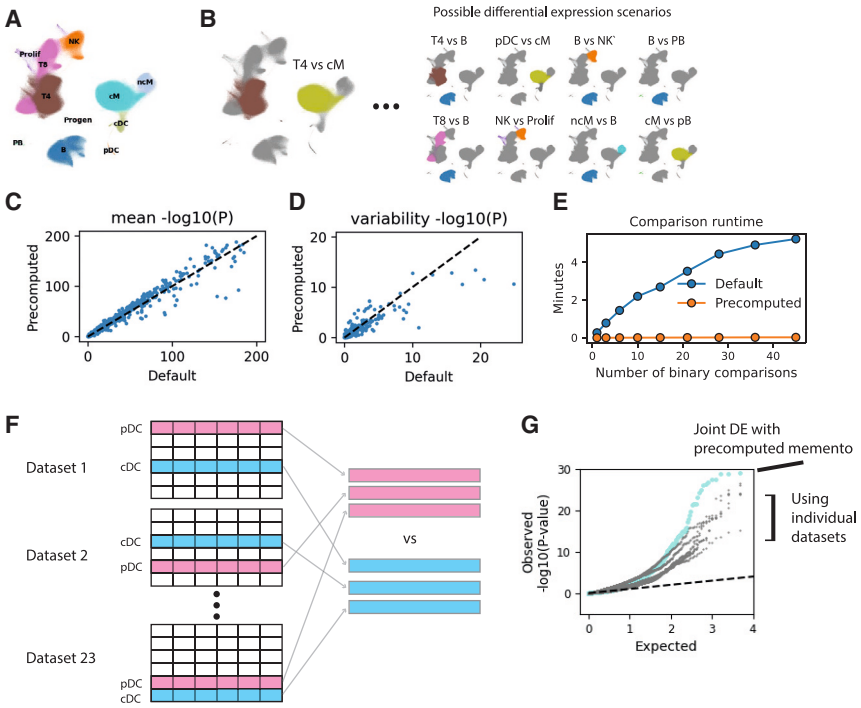


Figure 6. Extending Memento for near real-time differential expression analysis within CZI CELLxGENE Discover

(A) UMAP of the SLE PBMC dataset within CELLxGENE.

(B) Enumeration of different comparisons that can be made within and between groups of cells.

(C and D) 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 the 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 (gray).

See also Figure S6.

identify DM genes between conventional (cDC) and plasmacytoid dendritic cells (pDC). These cell types constitute 5.8% and 4.0%, respectively, of the scRNA-seq datasets of immune cells within the CELLxGENE Discover (Figure 6F). 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 with analyses of any single dataset (Figure 6G). These results highlight the efficiency of Memento's moment estimators and the adaptability of its bootstrap approach enable its effective application in expansive census repositories.

DISCUSSION

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,[12] disparate genetic perturbations induced by genome editing,[14,54] and individuals inheriting different alleles.[16–18] Initial observations that experimental and genetic perturbations predominantly induce subtle shifts in gene expression rather than unequivocal cell states have highlighted 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 scRNA-seq.

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: MoMs estimators modeling scRNA-seq via a hypergeometric sampling process and an efficient bootstrapping strategy to construct precise CIs around parameter estimates, exploiting the sparsity of scRNA-seq data. The utilization of MoMs estimators imparts 2-fold 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.[22] Second, our approach circumvents the need to repetitively compute the 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[33] and to design dimensionality reduction techniques for count data.[55] 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.[36] 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., scRNA-seq[30]), 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[56] and spatial transcriptomics data.[57] 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 MoMs estimators for hierarchical models universally contends with the challenge of establishing CIs via resampling, given that incorporating the sampling process into deriving analytical CIs 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. Because our hypothesis testing framework utilizes approximate bootstrapping, it should theoretically 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 front loading 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 to have 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 DM analysis. 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.

Limitations of the study

The current iteration of Memento presents a few limitations that offer opportunities for improvement in subsequent versions. Firstly, due to the bootstrapping method employed, Memento does not currently support the inclusion of cell-level covariates or continuous sample covariates. In both cases, one approach could be to group cells based on cell-level covariates or discretization of continuous covariates. Secondly, our method's capability for joint gene analyses is restricted to estimating and comparing gene correlations. Given that many biological pathways operate within lower-dimensional manifolds, future enhancements should enable comprehensive joint analyses involving more than two genes. Additionally, the flexibility inherent in the hypothesis testing framework of Memento should facilitate these adaptations seamlessly.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Chun Jimmie Ye (jimmie.ye@ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- scRNA-seq data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication. URLs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We acknowledge the use of artificial intelligence tools in the writing and editing processes of this manuscript. These tools assisted in organizing and refining the content, improving the clarity and coherence of the final document. The authors retain full responsibility for the intellectual content and accuracy of the research presented. C.J.Y. is supported by the NIH grants R01HG011239, R01AI136972, and R01AI045073 as well as by the Chan Zuckerberg Initiative and the Cancer Research Institute. He is also an investigator at the Chan Zuckerberg Biohub and the Arc Institute and a member of the Parker Institute for Cancer Immunotherapy (PICI). The research presented in this paper has also benefited from support through the JDRF and the NIH's Resource-based Center for the Advancement of Precision Medicine in Rheumatology.

AUTHOR CONTRIBUTIONS

M.C.K. and C.J.Y. conceived the project and wrote the paper. M.C.K. developed Memento and performed all analyses. D.S.L. and E.G. generated the HTEC data. A.L. and R.G. generated the Perturb-seq data. A.T. and P.E.G.-N. contributed to the implementation of Memento within CELLxGENE Discover. V.N., E.S., and A.M. provided valuable statistical and experimental feedback.

DECLARATION OF INTERESTS

C.J.Y. is founder for and holds equity in DropPrint Genomics (now ImmunAI) and Survey Genomics; a Scientific Advisory Board member for and holds equity in Related Sciences and ImmunAI; and a consultant for and holds equity in Maze Therapeutics. C.J.Y. has received research support from the Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, Arc Institute, Parker Institute for Cancer Immunotherapy, Genentech, BioLegend, ScaleBio, and Illumina. A.M. is a co-founder of Site Tx, Arsenal Biosciences, Spotlight Therapeutics, and Survey Genomics; serves on the boards of directors of Site Tx, Spotlight Therapeutics, and Survey Genomics; and is a member of the scientific advisory boards of Site Tx, Arsenal Biosciences, Cellanome, Spotlight Therapeutics, Survey Genomics, NewLimit, Amgen, and Tenaya. A.M. owns stock in Arsenal Biosciences, Site Tx, Cellanome, Spotlight Therapeutics, NewLimit, Survey Genomics, Tenaya, and Lightcast and has received fees from Site Tx, Arsenal Biosciences, Cellanome, Spotlight Therapeutics, NewLimit, Gilead, Pfizer, 23andMe, PACT Pharma, Juno Therapeutics, Tenaya, Lightcast, Trizell, Vertex, Merck, Amgen, Genentech, GLG, ClearView Healthcare, AlphaSights, Rupert Case Management, Bernstein, and ALDA. A.M. is an investor in and informal advisor to Offline Ventures and a client of EPIQ. The Marson laboratory has received research support from the Parker Institute for Cancer Immunotherapy, the Emerson Collective, Arc Institute, Juno Therapeutics, Epinomics, Sanofi, GlaxoSmithKline, Gilead, and Anthem and reagents from Genscript and Illumina.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Received: October 22, 2022
Revised: March 6, 2024
Accepted: September 26, 2024
Published: October 24, 2024

REFERENCES

1. McAdams, H.H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc. Natl. Acad. Sci. USA *94*, 814–819. https://doi.org/10.1073/pnas.94.3.814.
2. Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226. https://doi.org/10.1016/j.cell.2008.09.050.
3. Guo, J., and Zhou, X. (2015). Regulatory T cells turn pathogenic. Cell. Mol. Immunol. *12*, 525–532. https://doi.org/10.1038/cmi.2015.12.
4. Newman, J.R.S., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L., and Weissman, J.S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature *441*, 840–846. https://doi.org/10.1038/nature04785.
5. Raj, A., Rifkin, S.A., Andersen, E., and Van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913–918. https://doi.org/10.1038/nature08781.
6. Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167–173. https://doi.org/10.1038/nature09326.
7. Hansen, M.M.K., Desai, R.V., Simpson, M.L., and Weinberger, L.S. (2018). Cytoplasmic amplification of transcriptional noise generates substantial cell-to-cell variability. Cell Syst. *7*, 384–397.e6. https://doi.org/10.1016/j.cels.2018.08.002.
8. Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-time kinetics of gene activity in individual bacteria. Cell *123*, 1025–1036. https://doi.org/10.1016/j.cell.2005.09.031.
9. Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. Science *336*, 183–187. https://doi.org/10.1126/science.1216379.
10. Gupta, A., Martin-Rufino, J.D., Jones, T.R., Subramanian, V., Qiu, X., Grody, E.I., Bloemendal, A., Weng, C., Niu, S.Y., Min, K.H., et al. (2022). Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Proc. Natl. Acad. Sci. USA *119*, e2207392119. https://doi.org/10.1073/pnas.2207392119.
11. Li, K.-C. (2002). Genome-wide coexpression dynamics: theory and application. Proc. Natl. Acad. Sci. USA *99*, 16875–16880. https://doi.org/10.1073/pnas.252466999.
12. Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L., et al. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. Science *367*, 45–51. https://doi.org/10.1126/science.aax6234.
13. Datlinger, P., Rendeiro, A.F., Boenke, T., Krausgruber, T., Barreca, D., and Bock, C. (2019). Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing. Preprint at bioRxiv. https://doi.org/10.1101/2019.12.17.879304.
14. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038.
15. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., et al. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94. https://doi.org/10.1038/nbt.4042.
16. Van Der Wijst, M.G.P., Brugge, H., De Vries, D.H., Deelen, P., Swertz, M.A., and LifeLines Cohort Study; BIOS Consortium, and Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat. Genet. *50*, 493–497. https://doi.org/10.1038/s41588-018-0089-9.
17. Perez, R.K., Gordon, M.G., Subramaniam, M., Kim, M.C., Hartoularos, G.C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. Science *376*, eabf1970. https://doi.org/10.1126/science.abf1970.
18. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., et al. (2022). Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science *376*, eabf3041. https://doi.org/10.1126/science.abf3041.
19. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. https://doi.org/10.1038/s41467-021-25960-2.
20. Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255–261. https://doi.org/10.1038/nmeth.4612.
21. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. *21*, 31. https://doi.org/10.1186/s13059-020-1926-6.
22. Sarkar, A., and Stephens, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing

analysis. Nat. Genet. *53*, 770–777. https://doi.org/10.1038/s41588-021-00873-4.

23. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.
24. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. https://doi.org/10.1038/s41467-017-02554-5.
25. Korthauer, K.D., Chu, L.F., Newton, M.A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. *17*, 222. https://doi.org/10.1186/s13059-016-1077-y.
26. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.
27. Eling, N., Richard, A.C., Richardson, S., Marioni, J.C., and Vallejos, C.A. (2018). Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. Cell Syst. *7*, 284–294.e12.
28. McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., Hu, J.L., Murrow, L.M., Weissman, J.S., Werb, Z., et al. (2019). MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat. Methods *16*, 619–626. https://doi.org/10.1038/s41592-019-0433-8.
29. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224. https://doi.org/10.1186/s13059-018-1603-1.
30. Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science *357*, 661–667. https://doi.org/10.1126/science.aam8940.
31. Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N.R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. Proc. Natl. Acad. Sci. USA *115*, E6437–E6446. https://doi.org/10.1073/pnas.1721085115.
32. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.
33. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.
34. Grün, D., Kester, L., and Van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640. https://doi.org/10.1038/nmeth.2930.
35. Ghazanfar, S., Lin, Y., Su, X., Lin, D.M., Patrick, E., Han, Z.-G., Marioni, J.C., and Yang, J.Y.H. (2020). Investigating higher-order interactions in single-cell data with scHOT. Nat. Methods *17*, 799–806. https://doi.org/10.1038/s41592-020-0885-x.
36. Zhang, M.J., Ntranos, V., and Tse, D. (2020). Determining sequencing depth in a single-cell RNA-seq experiment. Nat. Commun. *11*, 774. https://doi.org/10.1038/s41467-020-14482-y.
37. Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. *6*, 171–179.e5. https://doi.org/10.1016/j.cels.2018.01.014.
38. Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., and Zhang, N.R. (2018). SAVER: gene expression recovery for single-cell RNA sequencing. Nat. Methods *15*, 539–542. https://doi.org/10.1038/s41592-018-0033-z.
39. Efron, B., and Tibshirani, R. (1994). An Introduction to the Bootstrap (Chapman & Hall), pp. 436.
40. Andreoletti, G., Lanata, C.M., Trupin, L., Paranjpe, I., Jain, T.S., Nititham, J., Taylor, K.E., Combes, A.J., Maliskova, L., Ye, C.J., et al. (2021). 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*, 488. https://doi.org/10.1038/s42003-021-02000-9.
41. Goel, R.R., Kotenko, S.V., and Kaplan, M.J. (2021). Interferon lambda in inflammation and autoimmune rheumatic diseases. Nat. Rev. Rheumatol. *17*, 349–362. https://doi.org/10.1038/s41584-021-00606-1.
42. Zhang, L., Bukreyev, A., Thompson, C.I., Watson, B., Peeples, M.E., Collins, P.L., and Pickles, R.J. (2005). Infection of ciliated cells by human parainfluenza virus type 3 in an in vitro model of human airway epithelium. J. Virol. *79*, 1113–1124. https://doi.org/10.1128/JVI.79.2.1113-1124.2005.
43. Wu, N.-H., Yang, W., Beineke, A., Dijkman, R., Matrosovich, M., Baumgärtner, W., Thiel, V., Valentin-Weigand, P., Meng, F., and Herrler, G. (2016). The differentiated airway epithelium infected by influenza viruses maintains the barrier function despite a dramatic loss of ciliated cells. Sci. Rep. *6*, 39668. https://doi.org/10.1038/srep39668.
44. Ravindra, N.G., Alfajaro, M.M., Gasque, V., Huston, N.C., Wan, H., Szigeti-Buck, K., Yasumoto, Y., Greaney, A.M., Habet, V., Chow, R.D., et al. (2021). 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*, e3001143. https://doi.org/10.1371/journal.pbio.3001143.
45. Gough, D.J., Messina, N.L., Clarke, C.J.P., Johnstone, R.W., and Levy, D.E. (2012). Constitutive Type I interferon modulates homeostatic balance through tonic signaling. Immunity *36*, 166–174. https://doi.org/10.1016/j.immuni.2012.01.011.
46. Bradley, K.C., Finsterbusch, K., Schnepf, D., Crotta, S., Llorian, M., Davidson, S., Fuchs, S.Y., Staeheli, P., and Wack, A. (2019). Microbiota-driven tonic interferon signals in lung stromal cells protect from influenza virus infection. Cell Rep. *28*, 245–256.e4. https://doi.org/10.1016/j.celrep.2019.05.105.
47. Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.E., Proserpio, V., Donati, G., et al. (2018). Gene expression variability across cells and species shapes innate immunity. Nature *563*, 197–202. https://doi.org/10.1038/s41586-018-0657-2.
48. Mostafavi, S., et al. (2016). Parsing the interferon transcriptional network and its disease associations in brief resource parsing the interferon transcriptional network and its disease associations. Cell *164*, 564–578.
49. Shifrut, E., Carnevale, J., Tobin, V., Roth, T.L., Woo, J.M., Bui, C.T., Li, P.J., Diolaiti, M.E., Ashworth, A., and Marson, A. (2018). Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. Cell *175*, 1958–1971.e15. https://doi.org/10.1016/j.cell.2018.10.024.
50. Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. *50*, 1140–1150. https://doi.org/10.1038/s41588-018-0156-2.
51. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.
52. Nathan, A., Asgari, S., Ishigaki, K., Valencia, C., Amariuta, T., Luo, Y., Beynor, J.I., Baglaenko, Y., Suliman, S., Price, A.L., et al. (2022). Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. Nature *606*, 120–128. https://doi.org/10.1038/s41586-022-04713-1.

53. CZI Single-Cell Biology Program (2023). CZ Cell × GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at bioRxiv. https://doi.org/10.1101/2023.10.30.563174.

54. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*, 2559–2575.e28. https://doi.org/10.1016/j.cell.2022.05.013.

55. Townes, F.W., Hicks, S.C., Aryee, M.J., and Irizarry, R.A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. *20*, 295. https://doi.org/10.1186/s13059-019-1861-6.

56. 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.

57. Tian, L., Chen, F., and Macosko, E.Z. (2023). The expanding vistas of spatial transcriptomics. Nat. Biotechnol. *41*, 773–782. https://doi.org/10.1038/s41587-022-01448-2.

58. De Jager, P.L., Hacohen, N., Mathis, D., Regev, A., Stranger, B.E., and Benoist, C. (2015). ImmVar project: Insights and design considerations for future studies of "healthy" immune variation. Semin Immunol *27*, 51–57. https://doi.org/10.1016/j.smim.2015.03.003.

59. Hill, A.J., McFaline-Figueroa, J.L., Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J., Jackson, D., Shendure, J., and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. Nat. Methods *15*, 271–274. https://doi.org/10.1038/nmeth.4604.

60. Fu, Y., Wu, P.H., Beane, T., Zamore, P.D., and Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. BMC Genomics *19*, 531. https://doi.org/10.1186/s12864-018-4933-1.

61. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

62. Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika *40*, 237–264. https://doi.org/10.1093/biomet/40.3-4.237.

63. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

64. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

65. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902. https://doi.org/10.1038/nbt.2931.

66. Lun, A.T.L., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. *17*, 75. https://doi.org/10.1186/s13059-016-0947-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human tracheal epithelial cells	Cells were harvested from deceased organ donors according to established protocols	PMID: 1616056
Primary human CD4+ T cells	Primary human CD4+ T cells were isolated from peripheral blood mononuclear cells (PBMCs) by magnetic negative selection using the EasySep Human CD4+ T Cell Isolation Kit	STEMCELL, Cat #17952
Critical commercial assays		
10X Chromium Single Cell v2	10X Genomics	PN-120237
Deposited data		
Single-cell RNA sequencing of genetically engineered CD4+ T cells	This paper	GEO: GSE274751
Single-cell RNA sequencing of interferon stimulated HTECs	This paper	GEO: GSE274751
Single-cell RNA sequencing of PBMCs	Perez et al.[17]	GEO: GSE174188

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human tracheal epithelial cells

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.

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,[58] 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.

METHOD DETAILS

Inteferon stimulation of HTECs

Starting on day 27, interferon stimulation (IFN-β: 10 ng/ml, IFN-α2: 10 ng/ml, IFN-γ: 10 ng/ml, IFN-λ2: 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 μ 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.

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. 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 in activated accessible chromatin regions. We took the union of the highly expressed regulators and accessible binding sites, for a total of 140 regulators (Figure 1B).

The backbone plasmid used to clone the CROP-Seq library was CROPseq-Guide-Puro, purchased from Addgene (Addgene. Plasmid #86708). We used two sgRNAs oligo sequences from the Brunello library 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.[13] Lentivirus was produced using the UCSF ViraCore.

SLICE experiment and sequencing

Primary human CD4+ T cells were isolated from peripheral blood mononuclear cells (PBMCs) by magnetic negative selection using the EasySep Human CD4+ 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 (98%).[59]

QUANTIFICATION AND STATISTICAL ANALYSIS

Modeling scRNA-seq as a hypergeometric sampling process

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,[60] noise stemming from imperfect transcript capture for RT and imperfect cDNA sampling during sequencing persists, culminating in the observed, attenuated distribution of counts.

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:

$$\mathbf{Z}_c \sim P_Z, \text{expressed transcript counts in cell } c$$

$$N_c = \mathbf{1}^T \mathbf{Z}_c, \text{total transcript count of cell } c$$

$$\mathbf{X}_c = \frac{Z_c}{N_c}, \text{normalized transcript counts in cell } c$$

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

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) (Figures S1A and S1B).[61]

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

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

In previous works,[36] the full hypergeometric treatment was simplified by a series of approximations, starting from the hypergeometric model to the Poisson model:

$$\mathbf{Y}_c \sim \text{MultiHG}(\mathbf{Z}_c, N_c, qN_c), \text{observed transcript counts}$$

$$\mathbf{Y}_c \sim \text{Multinomial}\left(\frac{\mathbf{Z}_c}{N_c}, qN_c\right), \text{observed transcript counts}$$

$$Y_{cg} \sim \text{Bn}\left(\frac{Z_{cg}}{N_c}, qN_c\right), \text{observed transcript counts}$$

$$Y_{cg} \sim \text{Poi}\left(\frac{Z_{cg}}{N_c} qN_c\right), \text{observed transcript counts}$$

$$Y_{cg} \sim \text{Poi}(qZ_{cg}), \text{observed transcript counts}$$

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 Zhang et al.[36] 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.

$$\begin{aligned}
\mathbb{E}\left[Y_{cg} \middle| X_{cg}, N, q\right] &= X_{cg}N_cq \\
\mathbb{E}\left[Y_{cg}^2 \middle| X_{cg}, N, q\right] &= X_{cg}^2N_c^2q_c^2 + X_{cg}q_cN_c \\
\mathbb{E}\left[Y_{cg_i}Y_{cg_j} \middle| X_{cg_i}, X_{cg_j}, N, q\right] &= \mathbb{E}\left[X_{cg_i}X_{cg_j}N_c^2q^2 \middle| X_{cg_i}, X_{cg_j}, N, q\right] = X_{cg_i}X_{cg_j}N_c^2q^2
\end{aligned} \quad \text{(Equation 2)}$$

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

$$\mathbb{E}\left[Y_{cg}^2 - Y_{cg} \middle| X_{cg}, N, q\right] = X_{cg}^2N_c^2q^2 \quad \text{(Equation 3)}$$

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

$$\widehat{\mu}_{g,Poi} = \widehat{\mathbb{E}}\left[X_{cg}\right] = \frac{1}{n_{cells}}\sum_c \frac{Y_{cg}}{N_c q}$$

$$\widehat{\sigma}^2_{g,Poi} = \widehat{\mathbb{E}}\left[X^2_{cg}\right] - \widehat{\mathbb{E}}\left[X_{cg}\right]^2 = \frac{1}{n_{cells}}\sum_c \frac{Y^2_{cg} - Y_{cg}}{N_c^2 q^2} - \left(\frac{1}{n_{cells}}\sum_c \frac{Y_{cg}}{N_c q}\right)^2 \quad \text{(Equation 4)}$$

$$\widehat{\sigma}_{g_i g_j,Poi} = \widehat{\mathbb{E}}\left[X_{cg_i}X_{cg_j}\right] - \widehat{\mathbb{E}}\left[X_{cg_i}\right]\widehat{\mathbb{E}}\left[X_{cg_j}\right] = \frac{1}{n_{cells}}\sum_c \frac{Y_{cg_i}Y_{cg_j}}{N_c^2 q^2} - \left(\frac{1}{n_{cells}}\sum_c \frac{Y_{cg_i}}{N_c q}\right)\left(\frac{1}{n_{cells}}\sum_c \frac{Y_{cg_j}}{N_c q}\right)$$

Now, let us consider the full multivariate hypergeometric model, $\mathbf{Y}_c \sim = \text{MultiHG}(\mathbf{X}_c N_c, N_c, qN_c)$. For a random vector $\mathbf{A} \sim MultiHG(\mathbf{K}, N, n)$, the moments of A are:

$$\begin{aligned}\mathbb{E}[A_i] &= n\frac{K_i}{N}\\ \mathbb{E}\left[A_i^2\right] &= n\frac{N-n}{N-1}\frac{K_i}{N}\left(1-\frac{K_i}{N}\right) + n^2\frac{K_i^2}{N^2}\\ \mathbb{E}[A_i A_j] &= -n\frac{N-n}{N-1}\frac{K_i K_j}{N^2} + n^2\frac{K_i K_j}{N^2}\end{aligned} \quad \text{(Equation 5)}$$

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

$$\begin{aligned}\mathbb{E}\left[Y_{cg}\middle|X_{cg}, N_c, q\right] &= qN_c\frac{X_{cg}N_c}{N_c}\\ &= X_{cg}N_c q\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[Y^2_{cg}\middle|X_{cg}, N_c, q\right] &= qN_c\frac{N_c - qN_c}{N_c - 1}\frac{X_{cg}N_c}{N_c}\left(1-\frac{X_{cg}N_c}{N_c}\right) + q^2N_c^2\frac{X^2_{cg}N_c^2}{N_c^2}\\ &\approx qN_c(1-q)X_{cg}(1-X_{cg}) + q^2N_c^2X^2_{cg}\\ &= X^2_{cg}N_c^2q^2 + X_{cg}qN_c(1-q) - X^2_{cg}qN_c(1-q)\\ &= X^2_{cg}\left(N_c^2q^2 - N_cq(1-q)\right) + X_{cg}N_cq(1-q)\end{aligned} \quad \text{(Equation 6)}$$

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

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

$$\mathbb{E}\left[Y^2_{cg} - (1-q)Y_{cg}\middle|X_{cg}, N, q_c\right] = X^2_{cg}\left(N_c^2q^2 - N_cq(1-q)\right) \quad \text{(Equation 7)}$$

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 approximations 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:

$$\widehat{\mu}_{g,HG} = \widehat{\mathbb{E}}\left[X_{cg}\right] = \frac{1}{n_{cells}}\sum_c \frac{Y_{cg}}{N_c q}$$

$$\begin{aligned}\widehat{\sigma}^2_{g,HG} &= \widehat{\mathbb{E}}\left[X^2_{cg}\right] - \widehat{\mathbb{E}}\left[X_{cg}\right]^2\\ &= \frac{1}{n_{cells}}\sum_c \frac{Y^2_{cg} - Y_{cg}(1-q)}{N_c^2q^2 - N_cq(1-q)} - \left(\frac{1}{n_{cells}}\sum_c \frac{Y_{cg}}{N_c q}\right)^2\\ &\approx \frac{1}{n_{cells}}\sum_c \frac{Y^2_{cg} - Y_{cg}(1-q)}{N_c^2q^2} - \left(\frac{1}{n_{cells}}\sum_c \frac{Y_{cg}}{N_c q}\right)^2\end{aligned} \quad \text{(Equation 8)}$$

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

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

$$\widehat{\mu}_{g,naive} = \widehat{\mathbb{E}}\left[X_{cg}\right] = \frac{1}{n_{cells}}\sum_{c}\frac{Y_{cg}}{N_c q}$$

$$\widehat{\sigma}^2_{g,naive} = \widehat{\mathbb{E}}\left[X^2_{cg}\right] - \widehat{\mathbb{E}}\left[X_{cg}\right]^2 = \frac{1}{n_{cells}}\sum_{c}\frac{Y^2_{cg}}{N_c^2 q^2} - \left(\frac{1}{n_{cells}}\sum_{c}\frac{Y_{cg}}{N_c q}\right)^2 \quad \text{(Equation 9)}$$

$$\widehat{\sigma}_{g_i g_j,naive} = \widehat{\mathbb{E}}\left[X_{cg_i}X_{cg_j}\right] - \widehat{\mathbb{E}}\left[X_{cg_i}\right]\widehat{\mathbb{E}}\left[X_{cg_j}\right] = \frac{1}{n_{cells}}\sum_{c}\frac{Y_{cg_i}Y_{cg_j}}{N_c^2 q^2} - \left(\frac{1}{n_{cells}}\sum_{c}\frac{Y_{cg_i}}{N_c q}\right)\left(\frac{1}{n_{cells}}\sum_{c}\frac{Y_{cg_j}}{N_c q}\right)$$

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:

$$\lim_{q\to 0}\widehat{\sigma}^2_{g,HG} = \widehat{\sigma}^2_{g,Poi}$$
$$\lim_{q\to 1}\widehat{\sigma}^2_{g,HG} = \widehat{\sigma}^2_{g,naive} \quad \text{(Equation 10)}$$

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.[62] 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:

$$\frac{1}{N}(r+1)\frac{n_{r+1}}{n_r} \quad \text{(Equation 11)}$$

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

$$r_g = \sum_{c} Y_{cg}, \text{count of gene } g \text{ in the sample}$$
$$n_r = \sum_{g} 1\left(r_g = r\right), \text{number of genes with count } r$$
$$r^*_g = \left(r_g + 1\right)\frac{n_{r_g+1}}{n_{r_g}} \quad \text{(Equation 12)}$$
$$\widehat{\mu}_{g,memento} = \widehat{\mathbb{E}}\left[X_{cg}\right] = \frac{1}{\sum_c N_c} r^*_g$$

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 γ_c so that

$N_c q_c = n_{umi}\gamma_c$. The simplest way of estimating γ_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.[63–65] By default, Memento uses 10% of the least variable genes. After gene set $G*$ is formed by trimming variable genes, we compute γ_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 δ 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 median $\left(\sum_{g \in G*} Y_{cg}\right)$ over cells c as the δ value.

It is important to note that there are more sophisticated normalization methods that exist in literature.[66] 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 ovedispersion parameters.[27] 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 processs.[39] 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 $x_1, x_2, x_3, \ldots x_N$, we can model the number of appearance of each observation as a multinomial distribution with Multinomial$\left(N, \frac{1}{N} \ldots \frac{1}{N}\right)$. If there are K unique counts with n_k cells each, we can re-write the resampling distribution as Multinomial$\left(N, \frac{n_1}{N} \ldots \frac{n_K}{N}\right)$.

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 (Figures S2D and S2E).

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 Δt 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 Δt, from Δt_1 to Δt_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 Δt from $\Delta t_1, \ldots, \Delta t_B$ to form $\Delta t_1^*, \ldots, \Delta t_B^*$, similar to the strategy laid out in Efron and Tibshirani.[39] We can then compute the achieved significance level (ASL) for that test as

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

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:

$$\begin{bmatrix} \ln \mu_1 \\ \ln \mu_2 \\ \vdots \\ \ln \mu_{2n-1} \\ \ln \mu_{2n} \end{bmatrix}, \begin{bmatrix} \ln \tilde{\sigma}_1 \\ \ln \tilde{\sigma}_2 \\ \vdots \\ \ln \tilde{\sigma}_{2n-1} \\ \ln \tilde{\sigma}_{2n} \end{bmatrix}, \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_{2n-1} \\ \rho_{2n} \end{bmatrix} \sim \beta \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_{2n-1} \\ W_{2n} \end{bmatrix} + \alpha \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad \text{(Equation 13)}$$

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

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

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

This framework can easily extended to incorporating many covariates, including batch variables and interactions between variables of interest by introducing additional covariates into the model in equation 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.[39] Through extensive simulations, we confirm that for the test statistics we consider in Memento, this procedure yields well-calibrated results under the null hypothesis (Figure S2F). 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 (Figure 5).

Pre-processing the rIFNB1 PBMC dataset

We used the original clustering and tSNE visualization of the rIFNB1 dataset[15] 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.[15]

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 Zhang et al.[36] 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%.[15]

Extracting mean and variance from scRNA-seq data for simulation

We used the PBMC dataset[15] 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.[27]

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 2A, 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.

For mean, naive, Memento with $q = 0$ (Poisson estimator), and BASiCS estimates are identical. BASiCS is not applicable for gene correlation.

In addition, to investigate how a gene's mean expression influences the accuracy of variance estimation, we compared the true simulated mean with the average error in variance estimation.

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.

In this simulation, differences in the mean, variability, and correlation were retained for 150 genes and removed for the remainder (see STAR Methods). For the differential mean simulation, we created the dataset with biological replicates to mimic multiplexed experimental designs.[17]

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 Δm 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:

$$
\begin{aligned}
m_{1,unstim} &= N(m_{unstim}, 0.25) \\
m_{2,unstim} &= N(m_{unstim}, 0.25) \\
m_{1,stim} &= m_{1,unstim} + \Delta m + N(0, 0.25) \\
m_{2,stim} &= m_{2,unstim} + \Delta m + N(0.25)
\end{aligned}
$$

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 Δv 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:

$$
\begin{aligned}
v_{1,unstim} &= N(v_{unstim}, 0.25) \\
v_{2,unstim} &= N(v_{unstim}, 0.25) \\
v_{1,stim} &= v_{1,unstim} + \Delta v + N(0, 0.25) \\
v_{2,stim} &= v_{2,unstim} + \Delta v + N(0.25)
\end{aligned}
$$

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 a 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 2C.

For mean, naive, Memento with $q = 0$ (Poisson estimator), and BASiCS estimates are identical. BASiCS is not applicable for gene correlation, and SCVI/SAVER does not provide direct estimates of gene variabilities.

Clustering the HTEC transcriptomes

We performed filtering, normalization, and clustering with the scanpy[32] suite of tools using the default values. Cell types were manually identified based on previously known marker genes for HTECs.[44]

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-β 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.

Estimation of cutting efficiency

The cutting efficiencies are estimated as the 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 4E, we first used a list of pairs of regulators 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 (4F).

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 sensitivty measurements from Gough et al. where the authors compared the expression of ISGs in IFNAR1-KO and WT macrophages.[45] The fold-change between those two groups were defined as the tonic sensitivity, which is the number we use in Figure 3D.

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 erythematosis (SLE) and healthy controls. We maintained the same cell type classifications used in that study.

To identify eQTLs using the psedubulk approach, we first created pseuobulks 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.[17] For Memento, we also performed the test separately for the two populations, using the same genotypes and covariates. We used the hierarchial resampling mode for Memento.

Enrichment of eQTLs in ATAC peaks

We used the same set of ATAC peaks used by Perez et al.[17] 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 5, we compared the eQTLs we discovered using the two approaches to the eQTLs reported by Yazar et al.[18] 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 MxN 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 6, we used the donor covariates as one-hot encoded variables as well as their interaction terms with the cell type one-hot encoded variable.

CellPress
OPEN ACCESS

Supplemental figures

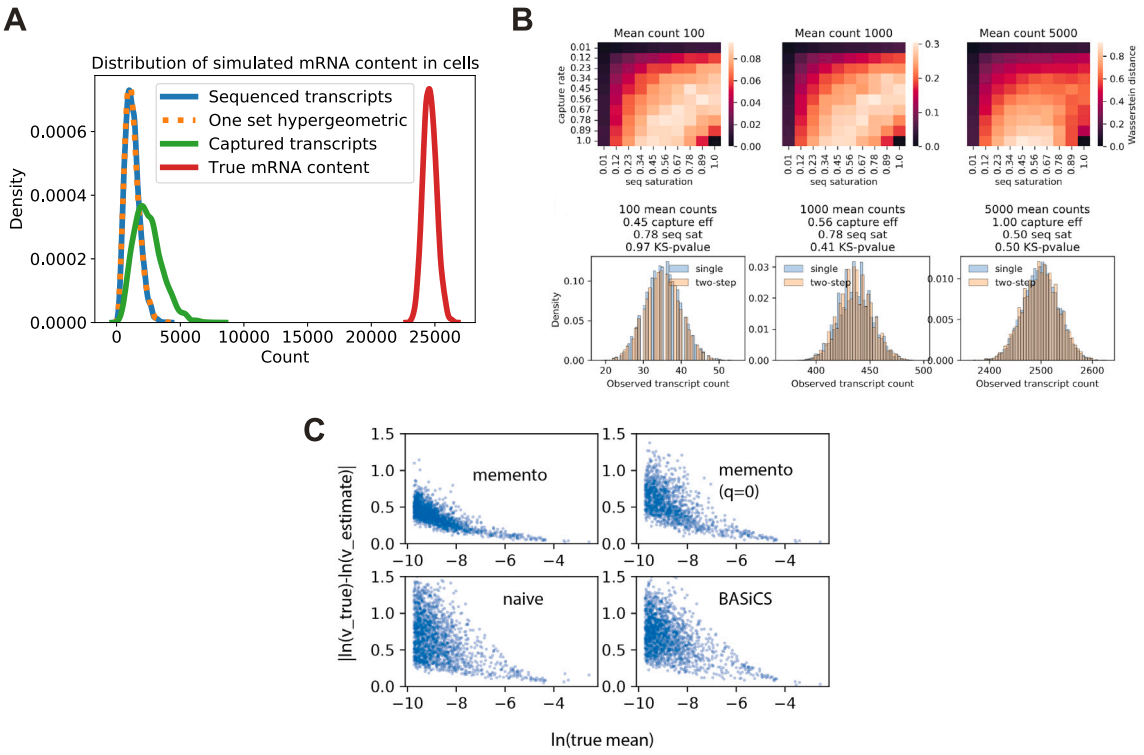
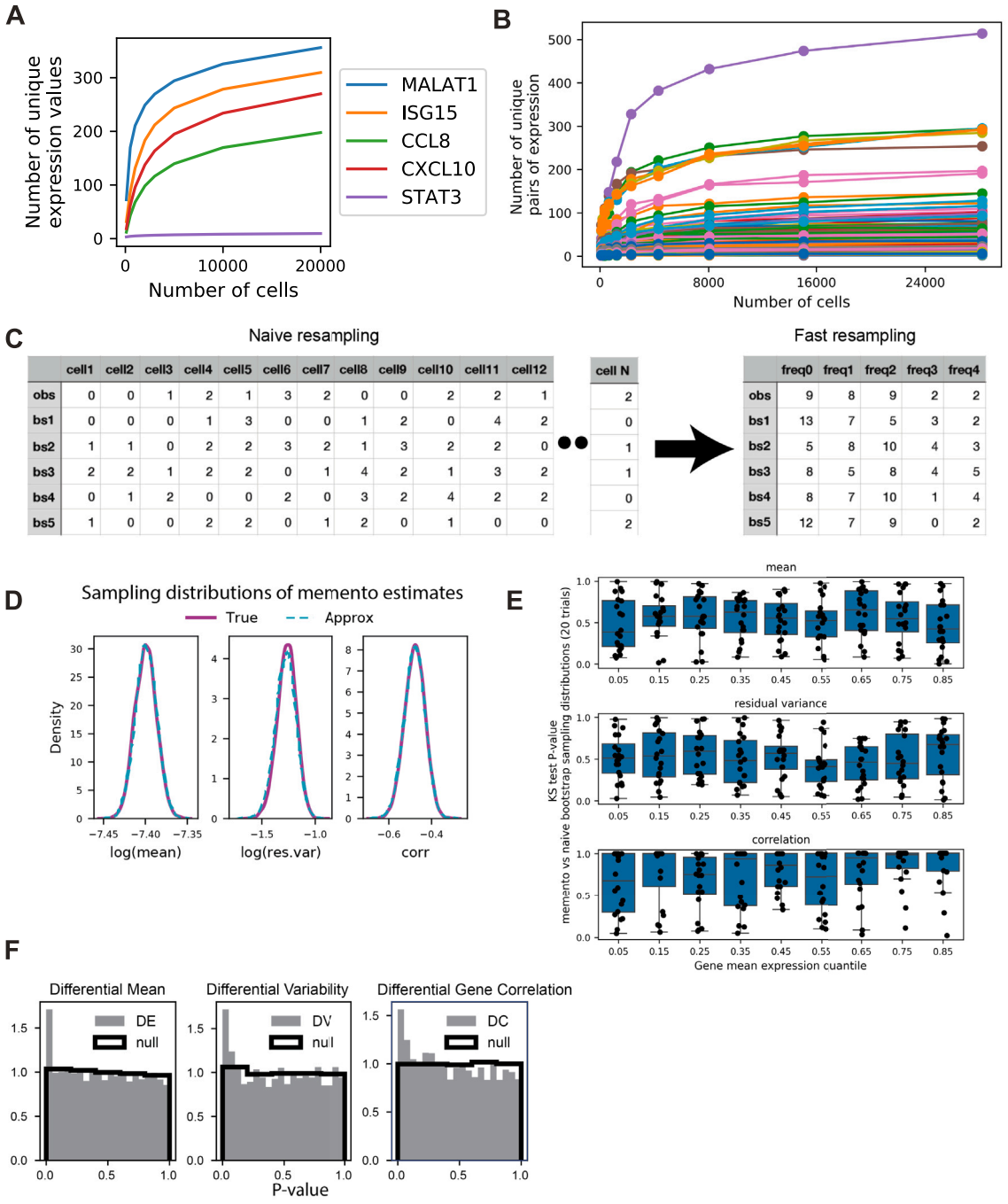


Figure S1. Validation of the hypergeometric sampling model, related to Figure 1

(A) Single step of hypergeometric sampling well approximates the compound sampling process from capture and sequencing.

(B) Characterizing the effect of approximating the two-step sampling in a single hypergeometric sampling step. Top: heatmap of Wasserstein distance between distributions resulting from various capture efficiency and sequencing saturation. Bottom: histogram of the two different sampling processes at a single capture efficiency and sequencing saturation (one of the tiles in A).

(C) 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.



	cell1	cell2	cell3	cell4	cell5	cell6	cell7	cell8	cell9	cell10	cell11	cell12	cell N
obs	0	0	1	2	1	3	2	0	0	2	2	1	2
bs1	0	0	0	1	3	0	0	1	2	0	4	2	0
bs2	1	1	0	2	2	3	2	1	3	2	2	0	1
bs3	2	2	1	2	2	0	1	4	2	1	3	2	1
bs4	0	1	2	0	0	2	0	3	2	4	2	2	0
bs5	1	0	0	2	2	0	1	2	0	1	0	0	2

	freq0	freq1	freq2	freq3	freq4
obs	9	8	9	2	2
bs1	13	7	5	3	2
bs2	5	8	10	4	3
bs3	8	5	8	4	5
bs4	8	7	10	1	4
bs5	12	7	9	0	2

Figure S2. Discrete nature of scRNA-seq data enables accurate, fast resampling, related to Figure 2

(A) Number of unique pairs of genes for randomly selected pairs in the IFN-B dataset.

(B) Number of unique pairs of genes for randomly selected pairs in the IFN-B dataset.

(C) Conceptual diagram for resampling frequencies rather than expression counts.

(D) Efficient bootstrap in Memento vs. full bootstrap. The sample distributions of log(mean), log(residual variance), and correlation for a representative gene and a pair of genes, respectively.

(E) Comprehensive comparison of Memento's bootstrap with the naive bootstrap. y axis is KS test *p* value from comparing the Memento vs. naive bootstrap distributions for the mean (top), residual variance (middle), and correlation (bottom). x axis is the gene's mean expression quantile.

(F) Representative example of *p* values computed from Memento.

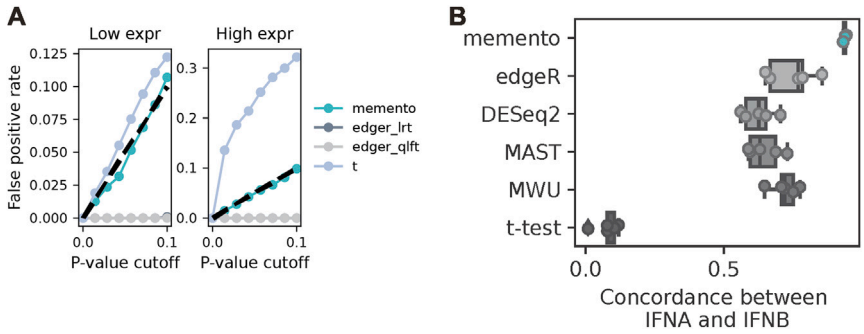


Figure S3. Further validation of Memento's differential mean approach, related to Figure 2

(A) False discovery rates in simulation for lowly and highly expressed genes.

(B) Concordance of DE genes between IFN-A and IFN-B across various methods.

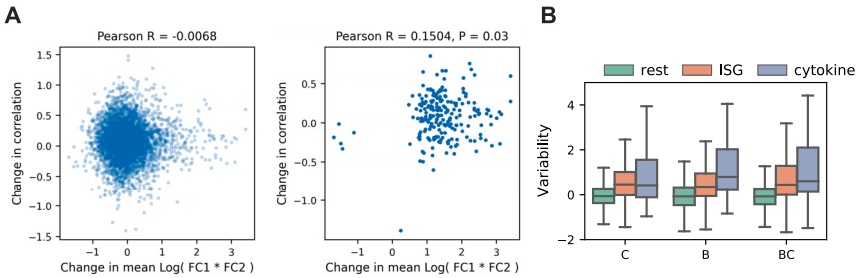


Figure S4. Patterns of gene co-expression and variability in HTEC data, related to Figure 3

(A) Change in correlation between two genes (y axis) vs. the product of the changes in mean (x axis) (left) and filtered by pairs with large mean change (right).
(B) Gene expression variability (y axis) for each class of genes across cell types.

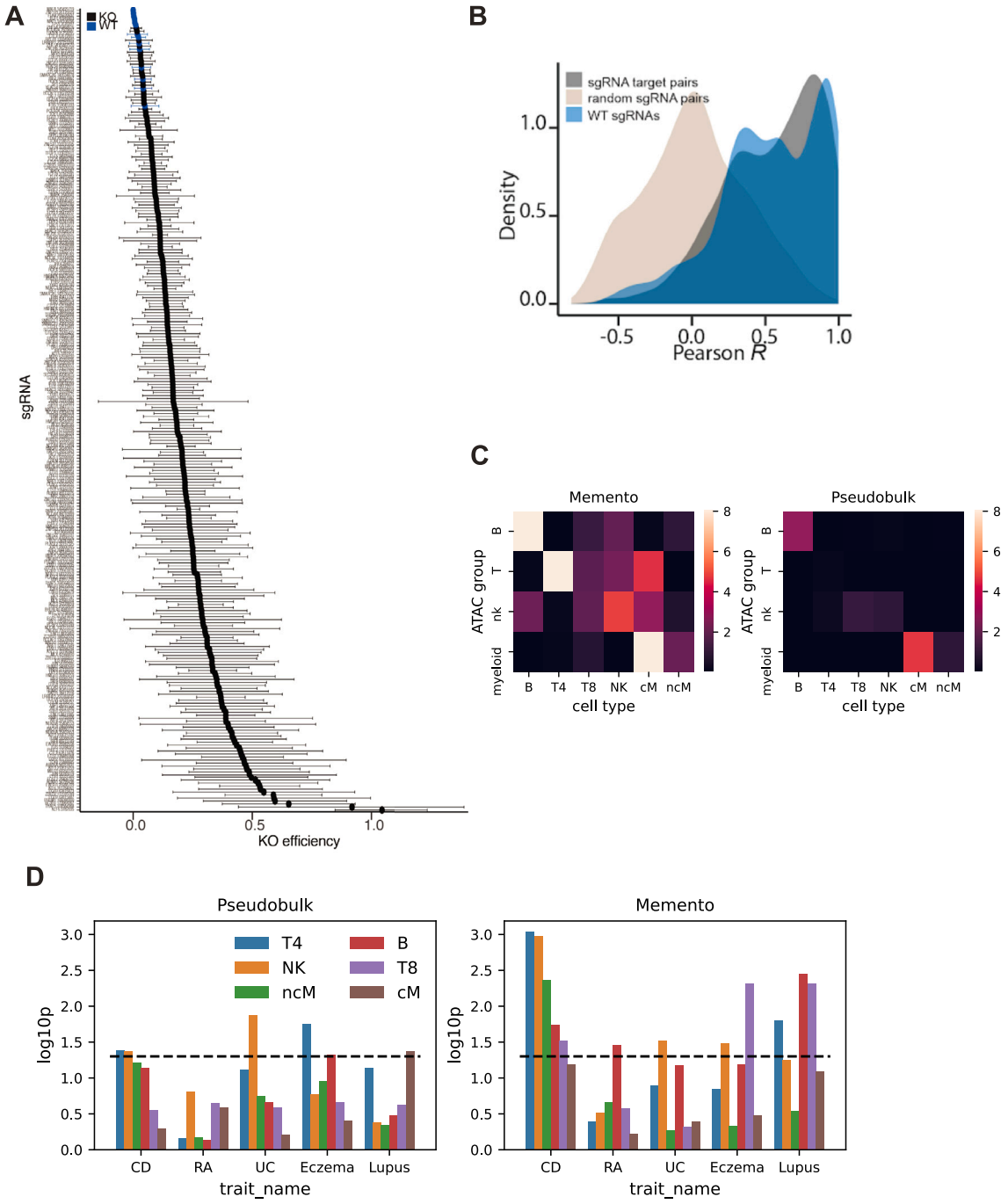


Figure S5. Additional experimental details and outcomes for analyses of genetic perturbations and natural genetic variability, related to Figure 4

(A) 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.

(B) Distribution of transcriptome correlations for WT guides, guides targeting the same gene, and random pairs of guides.

(C) Enrichment of eQTLs in cell-type-specific ATAC peaks (Europeans).

(D) LDSC-score regression enrichment for diseases using eGenes found via Memento and the pseudobulk method.

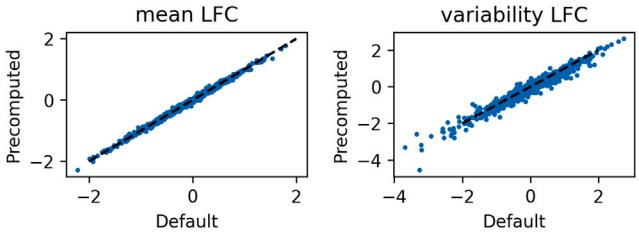


Figure S6. Comparison of effect sizes computed using full and precomputed versions of Memento, related to Figure 6