Title
Single-cell multi-omic analysis of immune cell development

Author
Steier, Zoe R

Publication Date
2021

Peer reviewed|Thesis/dissertation

Single-cell multi-omic analysis of immune cell development

By

Zoe R Steier

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy
with the University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Aaron Streets, Co-Chair
Professor Nir Yosef, Co-Chair
Professor Ellen Robey
Professor Jimmie Ye

Spring 2021

Abstract

Single-cell multi-omic analysis of immune cell development

By

Zoe R Steier

Doctor of Philosophy in Bioengineering

University of California, Berkeley

Professor Aaron Streets, Co-Chair, and Professor Nir Yosef, Co-Chair

The continuous differentiation and selection of T cells within the thymus is critical for the maintenance of mammalian adaptive immunity. Yet it is unclear precisely how thymocyte development and fate determination occur to produce T cells with different specified effector functions. Recent technological innovations in microfluidics and genomic sequencing have enabled high-throughput approaches for probing cell identities and development by measuring multiple molecular features in thousands of single cells. However, there has been a lack of computational methods capable of synthesizing this data to form a coherent view of cell identity. Here, I present a new method to analyze multi-omics data, describe how experimental and computational multi-omics analysis can be performed in practice, and apply these approaches to investigate T cell development.

First, I address the task of multi-omics data analysis. The paired measurement of RNA and surface proteins in single cells with CITE-seq is a promising approach to connect transcriptional variation with cell phenotypes and functions. However, combining these paired views into a unified representation of cell state is made challenging by the unique technical characteristics of each measurement. Here I present Total Variational Inference (totalVI), a deep generative model for end-to-end joint analysis of CITE-seq data that probabilistically represents the data as a composite of biological and technical factors including protein background and batch effects. To evaluate totalVI's performance, I profile immune cells from murine spleen and lymph nodes with CITE-seq, measuring over 100 surface proteins. I demonstrate that totalVI provides a cohesive solution for common analysis tasks like dimensionality reduction, the integration of datasets with different measured proteins, estimation of correlations between molecules, and differential expression testing.

Next, I present a guide for fellow researchers on how single-cell multi-omics analysis of RNA and proteins can be performed in practice. Despite the increasing availability of commercial experimental products and open-source software packages, there are many details and practical challenges that scientists must overcome in order to implement published methods in real-world settings across different biological contexts and experimental designs. Here I provide an overview of the experimental and computational pipelines for single-cell analysis of RNA and proteins. I then describe the practical steps necessary to complete these pipelines from collecting paired RNA

and protein data from single cells to preprocessing and filtering the sequencing data, running the totalVI model, and conducting downstream analysis. I also provide notes on common pitfalls and offer recommendations so that joint analysis of RNA and proteins can be applied widely to other biological systems.

Finally, I apply these methods for single-cell multi-omics analysis to investigate T cell development in the thymus. CD4 and CD8 T cells play a critical role in the mammalian immune system and understanding their fate decisions during development has broad clinical implications relevant to autoimmune diseases such as type 1 diabetes and to the production of cancer immunotherapies. While the development of CD4 and CD8 T cells within the thymus from the CD4+CD8+ stage has been widely studied as a classic model of a lineage determination, the developmental trajectory from immature thymocytes to mature T cells and the mechanism of lineage commitment remain unclear. To deconstruct this developmental process, I apply CITE-seq to simultaneously measure the transcriptome and over 100 surface proteins in thymocytes from wild-type and lineage-restricted mice. Using totalVI, I jointly analyze the paired measurements to build a comprehensive timeline of RNA and protein expression in the CD4 and CD8 lineages. Using lineage-restricted samples, I identify early differences that implicate the calcineurin-NFAT branch of the T cell receptor signaling pathway as a putative driver of lineage commitment. Employing drug perturbations in a neonatal thymic slice system, I validate the requirement of calcium signaling through NFAT for CD4, but not CD8, lineage commitment and shed light on the CD4/CD8 lineage commitment mechanism.

Table of Contents

Acknowledgments

My PhD has been full of intellectual exploration that has led me on a journey to a place that I never could have foreseen at the start. This journey has been guided and supported by many, without whom I would not have found my path. Here I would like to briefly acknowledge some of those who helped me along the way.

First, I would like to thank my co-advisors, Aaron Streets and Nir Yosef. I came to join their labs by an unconventional path, and I am very grateful that they took a chance on me. They gave me an incredible opportunity as their first co-advised student to work at the nexus between fields and to lead their work in new directions. I thank them both for generously sharing their knowledge with me, for teaching me how to ask good questions, and for supporting me through the ups and downs of the process. I am thankful not only for their fantastic guidance of my research, but also for guiding me as a scientist. Both Nir and Aaron believed in me and my work, and gave me the opportunity to learn multiple new fields, for which I am very grateful. From Aaron, I am thankful for the numerous enthusiastic discussions when we could get wrapped up in the science, temporarily forget about the passage of time, and share in the excitement of research. From Nir, I am thankful for the occasions when I could watch his thoughts crystallize, which many times clarified my own thinking and gently guided me towards more rigorous thinking myself. Both Nir and Aaron have a positivity and kindness about them and have gathered such wonderful lab members and collaborators that working as a part of their groups has truly been an exciting and fun experience.

I am very thankful to my lab mates in both the Streets and Yosef Labs for their continuous feedback and ideas through every stage of my projects. From the Streets Lab, I would especially like to thank Nick Altemose, Anushka Gupta, Gabriel Dorlhiac, and Annie Maslan for welcoming me and giving me their friendship and trust when I joined the group. Their positive and uplifting attitudes made me feel comfortable asking questions in a field that was new to me and gave me the confidence to chart new territory in my research. I am also grateful to Nick for leading me to Nir and Aaron, which made this entire dissertation possible. From the Yosef Lab, I am especially thankful for those I shared an office with – first David DeTomaso and Michael Cole, and then Matt Jones and Tal Ashuach. Finishing my PhD during a pandemic has made it especially apparent how much I enjoyed working with them and how valuable their advice was. I would also like to recognize David in particular for his patience and generosity in answering all of my questions as I caught up to speed in a new field. I don't think I would have progressed nearly as quickly without the chance to learn from David. I would also like to thank Jim Kaminski who in many ways served as a role model to me.

As a member of both the Streets and Yosef Labs and as my close collaborator, I owe special thanks to Adam Gayoso. By working with Adam, I had the incredible experience of being able to turn my idea into a reality. I am grateful to Adam for teaching me, for learning from me, for working through ideas together, and for all the time we worked side by side even when the pandemic forced us to do so remotely. In sharing these experiences with each other (including all of the bugs, the challenges, and the laughs), I felt like I had a true teammate.

I would like to thank Ellen Robey, who played many roles in my dissertation as a close collaborator, a dissertation committee member, and a qualifying exam committee member. I am grateful to Ellen for sharing her vast knowledge, perspective, and enthusiasm. I am also thankful for her guidance and support through every stage of my graduate work.

I was fortunate to be able to work with members of the Robey lab on multiple projects throughout graduate school. I would like to thank Silvia Ariotti, Lydia Lutes, Laura McIntyre, and Derek Bangs for everything they have taught me, for their willingness to take risks on my ideas, and for sharing their joy in science with me through our collaborations.

I would like to thank Jimmie Ye, who was a member of both my dissertation committee and qualifying exam committee, and Dave Schaffer and Ian Holmes, who were members of my qualifying exam committee. They each provided invaluable feedback and guidance. They also shared stimulating discussions that changed the way I thought about aspects of my work in ways that I have carried with me throughout the years.

Throughout these years in Berkeley, Zach Bleemer has been a continuous source of inspiration and critical discussions on this work and every subject. I am thankful for how he challenged me to think bigger about my work and about the world, and I am thankful for all that we shared.

I am incredibly grateful to Tiama Hamkins-Indik, Thomas Carey, and David Piech for their friendship, camaraderie, and support throughout graduate school. Our outdoor explorations that crisscrossed the Bay Area and spanned the West Coast brought me great joy. Knowing that we were all on this journey together has given me the strength to complete it.

Finally, I would like to thank my mom, Mandy, and my sister, Arielle, for believing in me from the very beginning. They cheered on every success and guided me through the toughest challenges I faced in graduate school. I am very grateful for their encouragement and endless support.

Chapter 1

Introduction

T cell development

T cells play a critical role in the adaptive immune system by recognizing and responding to foreign antigens displayed on the surface other cells. There are two main types of T cells that perform distinct functions: CD4 ("helper") T cells bind peptides on antigen presenting cells (APCs), activating the APC to produce antibodies (in the case of B cells) or to phagocytose a pathogen (in the case of macrophages), while CD8 ("cytotoxic") T cells bind peptides presented on target cells which they kill upon recognition. To maintain a properly functioning immune system, the development of T cells must be stringently regulated. Precursor cells are selected for the affinity of their T-cell receptor (TCR) for a specific antigen. Immature T cells that are unable to recognize an antigen with high enough affinity will die by neglect, while immature T cells that recognize self-peptides with too strong affinity will die by negative selection. Those with intermediate affinity will undergo positive selection to develop into mature T cells (Kurd et al., 2016). When this selection and maturation process is improperly regulated, diseases such as autoimmunity, infections, or cancer could result. To design therapies for these pathological cases, we must first have a firm understanding of how T cell development occurs in a healthy setting. Moreover, the knowledge to direct or control the development of CD4 and CD8 T cells could facilitate the engineering of immunotherapies, motivating the study of the mechanisms governing this developmental process.

T cells develop continuously in an organ called the thymus, where hematopoietic precursor cells first arrive from the bone marrow. Precursor cells known as thymocytes pass through multiple double negative developmental stages before beginning to express CD4 and CD8 surface proteins (double positive), which serve as co-receptors for TCR-antigen binding (Krueger et al., 2016). At this stage, cells expressing a functional TCR will bind to an antigen presented on either an MHCI or MHCII molecule of an APC, causing a signaling cascade that results in the maturation of the thymocyte into a single positive CD8 or CD4 T cell in which either CD4 or CD8, respectively, is no longer expressed (Kurd et al., 2016). This maturation process takes place over the course of two to three days during which the thymocytes migrate from the cortex to the medulla, with mature T cells eventually being transported out into circulation (Sinclair et al., 2013).

To understand the biology of mammalian thymocyte development, the mouse thymus serves as an excellent model system and has provided great insight on T cell selection and maturation. Studies in mice led to the discovery the genes of the master regulators *Zbtb7b* (encoding THPOK) for CD4 fate and *Runx3* for CD8 fate (He et al., 2005; Sun et al., 2005; Woolf et al., 2003). Once activated, these transcription factors drive the transcription of genes that promote one cell fate and inhibit the other (Kurd et al., 2016). Additional studies have demonstrated that the signaling pathways immediately downstream of the TCR rely on many of the same molecules in these two cell types.

It is still unknown how the binding of an antigen by the TCR of immature thymocytes plays a role in producing signals that result in the maturation into two distinct cell types (Vacchio et al., 2016). This exemplifies a fundamental problem in biology: how does a developing cell decide its fate? The CD4/CD8 fate decision is a classic model of a branching, irreversible lineage commitment. Yet, it remains unclear precisely how CD4 or CD8 fate is determined and how thymocytes progress through phenotypic changes to produce mature CD4 and CD8 T cells. In the following sections, I explore state-of-the-art methods that could be used to investigate these questions with unprecedented resolution. I apply these methods in Chapter 4 to conduct an investigation of thymocyte development.

Characterizing cell identity

When studying a developmental process at the cellular level, there are multiple aspects of a cell's identity that could be considered. We could describe a cell's identity as existing along a continuum of phenotypes that progress over time and space (Wagner et al., 2016). This phenotype could include factors such as the cell's physical morphology and its molecular composition. While imaging can provide information about the physical form, spatial position, and environment of the cell, it offers few clues to how or why a cell transitions from one developmental stage to another. On the molecular level, the unique set of proteins contained in a cell can characterize a cell's state and reflect the molecular mechanisms at play in a developmental process. Particularly in the field of immunology, surface proteins are commonly used to define cell phenotypes as they play a critical role in immune cell function and can be used to distinguish major cell types. However, the gold-standard techniques of flow cytometry or fluorescence-activated cell sorting (FACS) are limited to measuring 18 proteins on a single cell due to spectral overlap of fluorescent antibodies that are necessary to uniquely detect the proteins of interest (Papalexi and Satija, 2017). These methods are not capable of comprehensive characterization of the proteome, which is particularly challenging due to the lack of intracellular measurements of proteins such as transcription factors that regulate gene expression. They are also ill-suited to uncover new factors that play a role in development because the proteins of interest must be identified in advance in order to select the relevant antibodies. Due to the challenges in measuring proteins, RNA often serves as a proxy, as it encodes the information necessary to produce proteins but can be measured more directly through sequencing. The measurement of all RNA molecules in a cell would quantify which genes of the genome are being expressed, and thus represents a type of comprehensive characterization of cell identity.

From bulk genome measurements to high-throughput, single-cell multi-omics

The development of high-throughput RNA sequencing (RNA-seq) methods that quantify the complete set of transcripts in a sample was a revolutionary advance (Wang et al., 2009). RNA-seq has some clear advantages overs techniques like flow cytometry that are limited in the number of measurements that can be made and biased by the measurement of pre-selected molecules. However, the original RNA-seq methods were designed to be performed on samples of cells in bulk due to limited sensitivity to detect small numbers of RNA molecules. Because these bulk methods produce average measurements across a population of cells, they are inadequate for

observing heterogeneity across a continuous developmental process. Technological innovations to improve sensitivity paved the way for single-cell RNA sequencing (scRNA-seq), where RNA sequencing reactions are performed on individually isolated cells (Tang et al., 2009; Picelli et al., 2013). Subsequent advances in microfluidics increased the sensitivity and the number of single cells that can be measured (Streets et al., 2014, Wu et al., 2017). In particular, throughput was massively increased by methods that isolate thousands of single cells into nanoliter droplets, labeling them with DNA barcodes, and pooling them for sequencing reactions (Macosko et al., 2015; Klein et al., 2015). Such methods provided the sensitivity and scale to embark on large-scale efforts to identify and profile every cell type in the human body through the Human Cell Atlas project (Regev et al., 2017).

While scRNA-seq can achieve whole-transcriptome resolution, it lacks the functional and phenotypic information contained in the proteome, which has long been the gold-standard definition of cell types in immunology. Recently, the field has pushed towards developing multi-omics technologies in which multiple molecular components can be measured in the same single cell. Approaches such as CITE-seq and REAP-seq adapted droplet-based scRNA-seq methods to simultaneously measure the transcriptome and surface proteins (Stoeckius et al., 2017; Peterson et al., 2017). These methods appeared highly promising for the study of immune cells by providing measurements of two uniquely valuable modalities for cell type characterization. However, the data produced by these experiments posed major challenges in analysis.

Challenges in multi-omics analysis

At the advent of multi-omics methods for paired measurements of RNA and proteins, it was not at all clear how the data should be analyzed. Early studies using CITE-seq and REAP-seq tended to apply standard workflows to analyze the RNA portion of the data while using protein data to validate and aid the interpretation of the RNA analysis post-hoc. These approaches resulted in analyses that were not only biased towards one modality, but also contained bias due to technical factors in the protein measurement that were not adequately addressed. To fully leverage the multiple molecular views of a cell provided by multi-omic measurements, an analysis method should take both modalities into account when performing all stages of analysis including, for example, determining similarities between cells, grouping cells into types, and quantifying differences in molecular profiles across groups. More than performing analysis on both modalities in parallel and combining their interpretations, a joint analysis would take advantage of the knowledge that a paired set of measurements originated from the same cell and combine the views of a cell gained by each modality in an unbiased way.

To combine the data from the paired RNA and protein measurements, any analysis must first address the distinct sources of technical bias and noise in each modality. For the RNA data, technical aspects such as batch effects and sequencing library size have previously been addressed by computational methods such as scVI (Lopez et al., 2018), which uses deep learning to probabilistically model the biological and technical components of scRNA-seq data. However, the protein data has distinct sources of noise from RNA due to the different nature of the measurement:

rather than quantifying the molecules directly as in RNA sequencing, proteins cannot be directly sequenced, and thus are quantified via the sequencing of DNA barcodes that are conjugated to protein-specific antibodies. This results in large amounts of background in the protein measurement due to non-specifically bound or ambient antibodies. In addition, unlike the unbiased measurement of the whole transcriptome, the protein measurements rely on selections of antibodies that could vary across experiments, presenting a challenge in integrating datasets containing different measured proteins. In Chapter 2, I develop a computational framework called totalVI for multi-omic data analysis that relies on a detailed understanding of the experimental data-generating process to address its technical limitations and to jointly model the paired measurements (Gayoso et al., 2021).

Multi-omics analysis in practice

For researchers who want to use multi-omics analysis to answer a particular biological question, there are two major steps that must be completed: data collection and computational analysis. Especially for new researchers, both of these tasks can be daunting. Even with a firm understanding of the theoretical underpinnings of the experimental or computational methods, numerous questions tend to arise when performing these analyses in practice. On the experimental side, published methods often demonstrate proofs-of-concept in simple biological settings (e.g., healthy peripheral blood mononuclear cells), but have little guidance for how methods should be applied in other contexts, how much tolerance there is for protocol modifications, and what the tradeoffs are when making decisions that affect data quality, time, and cost. For instance, how many cells are needed, how deeply should RNA and protein libraries be sequenced, and what are places to troubleshoot when data quality is low? On the computational side, researchers without computational expertise often encounter basic questions related to performing a joint analysis: what are the required inputs, what are the outputs, and how should they be interpreted? Even for more experienced researchers, aspects unique to each dataset often raise questions: how should quality control filters be set and what hyperparameters should be used? To address these questions, I share in Chapter 3 a guide through the steps and decisions that must be made when conducting single-cell multi-omics analysis with RNA and proteins.

Scope of the dissertation

In Chapter 2, I present totalVI, a computational framework that probabilistically models paired RNA and protein data from single cells to enable joint analysis of both modalities. In Chapter 3, I provide a guide to the experimental and computational pipelines for conducting multi-omics analysis of RNA and protein data from single cells. In Chapter 4, I apply totalVI and the methods described in Chapter 3 to investigate thymocyte development, uncovering a pathway through which CD4 or CD8 T cell fate is determined. These works have opened the door to multiple future directions. In Chapter 5, I describe a selection of the ongoing works and promising next steps for future research building upon the work presented here.

References

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, *18*(3), 272–282. https://doi.org/10.1038/s41592-020-01050-x

He, X., He, X., Dave, V. P., Zhang, Y., Hua, X., Nicolas, E., … Kappes, D. J. (2005, February 24). The zinc finger transcription factor Th-POK regulates CD4 versus CD8 T-cell lineage commitment. *Nature*. Nature Publishing Group. https://doi.org/10.1038/nature03338

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Krueger, A., Ziętara, N., & Łyszkiewicz, M. (2017, February 1). T Cell Development by the Numbers. *Trends in Immunology*. Elsevier Ltd. https://doi.org/10.1016/j.it.2016.10.007

Kurd, N., & Robey, E. A. (2016). T-cell selection in the thymus: a spatial and temporal perspective. *Immunological Reviews*, *271*(1), 114–126. https://doi.org/10.1111/imr.12398

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Papalexi, E., & Satija, R. (2017). Single-cell RNA sequencing to explore immune cell heterogeneity. https://doi.org/10.1038/nri.2017.76

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., … Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*(10), 936–939. https://doi.org/10.1038/nbt.3973

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1100. https://doi.org/10.1038/nmeth.2639

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., … Yosef, N. (2017). The human cell atlas. *ELife*, *6*. https://doi.org/10.7554/eLife.27041

Sinclair, C., Bains, I., Yates, A. J., & Seddon, B. (2013). Asymmetric thymocyte death underlies the CD4:CD8 T-cell ratio in the adaptive immune system. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(31), E2905–E2914. https://doi.org/10.1073/pnas.1304859110

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., … Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868. https://doi.org/10.1038/nmeth.4380

Streets, A. M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., … Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(19), 7048–7053. https://doi.org/10.1073/pnas.1402030111

Sun, G., Liu, X., Mercado, P., Jenkinson, S. R., Kypriotou, M., Feigenbaum, L., … Bosselut, R. (2005). The zinc finger protein cKrox directs CD4 lineage differentiation during intrathymic T cell positive selection. *Nature Immunology*, *6*(4), 373–381. https://doi.org/10.1038/ni1183

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Vacchio, M. S., & Bosselut, R. (2021). Function Transcriptional Circuitry To Control Their Lineage Commitment + CD8 − + in the Thymus: How T Cells Recycle the CD4 What Happens in the Thymus Does Not Stay. *J Immunol References*, *196*, 4848–4856. https://doi.org/10.4049/jimmunol.1600415

Wagner, A., Regev, A., & Yosef, N. (2016, November 8). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*. Nature Publishing Group. https://doi.org/10.1038/nbt.3711

Wang, Z., Gerstein, M., & Snyder, M. (2009, January). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. NIH Public Access. https://doi.org/10.1038/nrg2484

Woolf, E., Xiao, C., Fainaru, O., Lotem, J., Rosen, D., Negreanu, V., … Groner, Y. (2003). Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 7731–7736. https://doi.org/10.1073/pnas.1232420100

Chapter 2

Joint probabilistic modeling of single-cell multi-omic data with totalVI

Adam Gayoso[*], Zoë Steier[*], Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets†, and Nir Yosef†

* These authors contributed equally.
† Corresponding authors

From Gayoso, A.*, Steier, Z.*, Lopez, R. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18,** 272–282 (2021). https://doi.org/10.1038/s41592-020-01050-x. Reprinted with permission from Springer Nature.

Abstract
The paired measurement of RNA and surface proteins in single cells with CITE-seq is a promising approach to connect transcriptional variation with cell phenotypes and functions. However, combining these paired views into a unified representation of cell state is made challenging by the unique technical characteristics of each measurement. Here we present Total Variational Inference (totalVI; https://scvi-tools.org), a framework for end-to-end joint analysis of CITE-seq data that probabilistically represents the data as a composite of biological and technical factors including protein background and batch effects. To evaluate totalVI's performance, we profiled immune cells from murine spleen and lymph nodes with CITE-seq, measuring over 100 surface proteins. We demonstrate that totalVI provides a cohesive solution for common analysis tasks like dimensionality reduction, the integration of datasets with different measured proteins, estimation of correlations between molecules, and differential expression testing.

Introduction
The advance of technologies for quantitative, high-throughput measurement of the molecular composition of single cells is continuously expanding our understanding of cell ontology, state, and function [1-3]. A growing body of single-cell multi-omic techniques now offers the ability to further refine our definitions of cellular identity by providing multiple views of molecular state [4, 5]. By extending single-cell RNA-sequencing (scRNA-seq) to simultaneously measure the abundance of proteins on the cell surface, CITE-seq [6,7] presents the opportunity to connect the information that can be gleaned from the transcriptome [8, 9] to the functional information contained in proteins [10, 11]. Such experimental tools necessitate computational tools to synthesize these high-dimensional views.

Recent studies have analyzed CITE-seq data using standard workflows for one modality (often RNA) to cluster cells while contextualizing these results using information from the other modality post-hoc [12-14]. This sequential approach biases the analysis to one modality and becomes increasingly inefficient as CITE-seq measurements expand to hundreds of proteins. A joint analysis that combines these two cellular views in an unbiased manner can harness the strengths of each modality and streamline data analysis. However, combining RNA and protein information to define a single representation of cell state poses several challenges. First, the RNA and protein data have unique sources of technical bias and noise. While the technical aspects of the RNA data have been addressed by a flourishing body of computational methods [15-18], the protein data present distinct technical bias such as background due to ambient or non-specifically bound antibodies. Second, as large-scale community efforts such as the Human Cell Atlas (HCA) [8] begin to include CITE-seq datasets, the need arises for scalable computational methods that can integrate datasets with different measured proteins.

Here, we present totalVI (Total Variational Inference), a deep generative model that enables multifaceted analysis of CITE-seq data and addresses these challenges. totalVI learns a joint probabilistic representation of the paired measurements that accounts for the distinct noise and technical biases of each modality, as well as batch effects. For RNA, totalVI uses a modeling strategy similar to our previous work (scVI; [15]). For proteins, totalVI introduces a new model that separates the protein signal into background and foreground components, which enables background correction. The probabilistic representations learned by totalVI are built on a joint low-dimensional representation of the RNA and protein data that is derived using neural networks.

totalVI can be used for disparate analysis tasks including joint dimensionality reduction, dataset integration (with and without missing proteins), protein background correction, estimation of correlations between genes and/or proteins, and differential expression testing. To highlight this functionality, we performed CITE-seq on murine spleen and lymph nodes, measuring up to 208 proteins. We used these data, along with public datasets, to evaluate totalVI's performance across these tasks.

Results

The totalVI model

totalVI uses a probabilistic latent variable model [19] to represent the uncertainty in the observed RNA and protein counts from a CITE-seq experiment as a composite of biological and technical sources of variation. The input to totalVI consists of the matrices of RNA and protein unique molecular identifier (UMI) counts (Fig. 1a). Categorical covariates such as experimental batch or donor are optional inputs used for integrating datasets and referred to henceforth as "batch". Input datasets can have different antibody panels, and a subset can be scRNA-seq datasets (i.e., without proteins).

The output of totalVI consists of two components that can be used for downstream analysis (Fig. 1b). The first component encodes each cell as a distribution in a low-dimensional latent space (20 dimensions throughout; Supplementary Note 1) that represents the information contained in both the RNA and protein data (Supplementary Note 2), while controlling for their respective noise properties and batch effects. The second component provides a way to estimate the parameters of the distributions that underlie the observed RNA and protein measurements (i.e., likelihoods) given a cell's latent representation. These distributions explicitly account for nuisance factors in the observed data such as sequencing depth, protein background, and batch effects (Supplementary Note 3). Both components use neural networks to specify distributions.

totalVI optimizes the parameters of both of its components simultaneously using the variational autoencoder (VAE) framework [20]. Accordingly, totalVI uses highly efficient techniques for stochastic optimization that make it appropriate for the scale of CITE-seq data. Following optimization, totalVI's components are used for downstream analysis. The latent cell representations can be used as input to methods that stratify cells like clustering, visualization, or pseudotime inference algorithms, thus allowing these methods to leverage both protein and RNA information. Other downstream tasks specific to genes and proteins, like differential expression, are linked to the likelihood parameters from the second component of totalVI. Finally, by constricting the latent space to the standard simplex, the dimensions of the latent space can be related to the expression of genes and proteins with archetypal analysis [21], adding an alternative way to investigate global and local patterns of variation in the data. A detailed specification of the model along with further description of the quantities used in downstream tasks is in Methods.

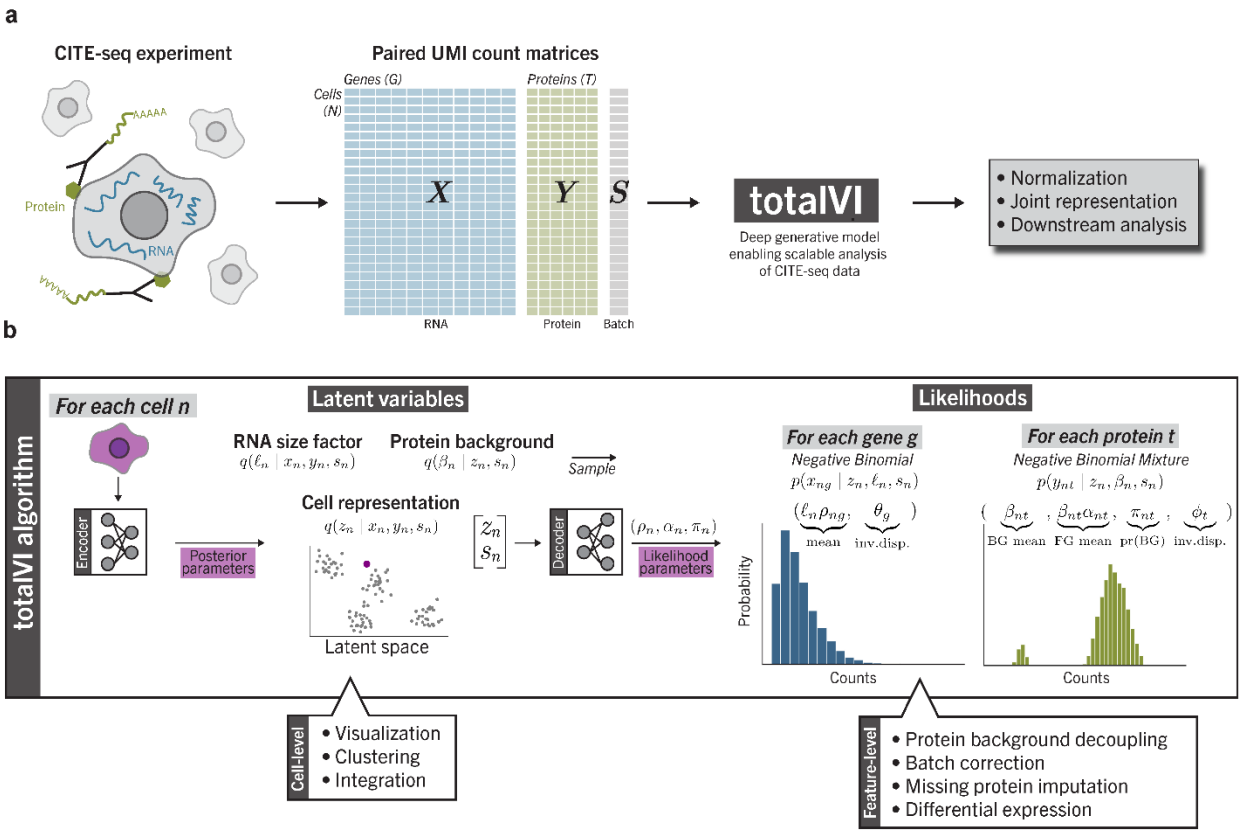


Figure 1: Schematic of a CITE-seq data analysis pipeline with totalVI. a, A CITE-seq experiment simultaneously measures RNA and surface proteins molecules in single cells, producing paired count matrices for RNA and proteins. These matrices, along with an optional matrix containing sample-level categorical covariates (batch), are the input to totalVI, which concomitantly normalizes the data and learns a joint representation of the data that is suitable for downstream analysis tasks. **b**, Schematic of totalVI model. The RNA counts, protein counts, and batch for each cell n are jointly transformed by an encoder neural network into the parameters of the posterior distributions for z_n, a low-dimensional representation of cell state, β_n, the protein background mean indexed by protein, and ℓ_n, an RNA size factor. The posterior mean of z_n, which we refer to as the latent representation, is corrected for batch effects and can be used as input to clustering and visualization algorithms. Next, a decoder neural network maps samples from the posterior distribution of z_n, along with the batch, s_n, to parameters of a negative binomial distribution for each gene and a negative binomial mixture for each protein, which contains a foreground (FG) and background (BG) component (Methods). These parameters are used for feature-level analyses.

CITE-seq profiling of murine spleen and lymph nodes

We conducted a series of CITE-seq experiments that were designed to test the performance of totalVI on a variety of tasks. As a case study, we profiled murine spleen and lymph nodes, which contain heterogeneous immune cell populations that are well-characterized by surface protein markers. Cells were collected from two wild-type mice that were processed on separate days to serve as biological replicates (Methods). In each experimental run, cells from one mouse were stained with two different panels of barcoded antibodies containing either 111 or 208 antibodies, of which the 111 antibodies were a subset (Supplementary Data). Spleen and lymph node cells stained separately with the same antibody panel were combined using hashtag antibodies [22]. We

refer to the four resulting spleen/lymph node datasets by their panel and experimental day (experimental design in Supplementary Table 1), After pre-processing and filtering, these datasets contained a total of 32,648 cells (Methods).

totalVI fits CITE-seq data well and is scalable

The usefulness of probabilistic models like totalVI depends on how well they fit the observed data. Furthermore, they should generalize to unobserved data (i.e., not overfit) and scale to a realistic range of input sizes. To verify that totalVI satisfies these prerequisites, we benchmarked it against factor analysis (FA), which can be viewed as a linear-Gaussian baseline, scHPF [16], which performs a Poisson matrix factorization via a hierarchical Bayesian model, and scVI [15], which was restricted to the RNA portion of the data. We expected the performance of totalVI and scVI to be comparable on the RNA data, as they share similar architectures. Our evaluation relied on fitting the models to several CITE-seq datasets spanning different species and tissues, including peripheral blood mononuclear cells (PBMC10k) [23] and mucosa-associated lymphoid tissue (MALT) [24] from humans, and our murine spleen and lymph node data (SLN111-D1).

We first estimated how well each model fit data that was available to it during training using posterior predictive checks (PPC) [16, 25]. To conduct PPCs, we generated replicated datasets (i.e., posterior predictive samples) by sampling from the fitted model (Methods). We then assessed how well these replicated datasets maintained the properties of the observed data with two metrics. First, we measured the similarity between the coefficient of variation (CV) per gene and protein of the replicated data to the observed CVs, thus evaluating how well the mean-variance relationship of the data is preserved. Second, we compared the replicated and raw data at the gene and protein level using the Mann-Whitney U statistic, which measures the extent to which the replicated and raw data come from the same distribution. totalVI had superior performance on both metrics (Extended Data Fig. 1a, b).

We then evaluated how well each model generalizes to cells that were not available during training by generating replicated datasets conditioned on the held-out cells and computing two opposing metrics of predictive performance. First, we assessed how well the average replicated data set matched the observed held-out data by mean absolute error. Second, we quantified how well the interval of values from replicated data sets covered the observed held-out data values (calibration error [26]). These two metrics were computed separately for genes and proteins. On the held-out protein data, totalVI outperformed FA in both the mean absolute error and calibration error metrics. Comparing totalVI to scHPF revealed a tradeoff between calibration and held-out error for both the RNA and protein data. On the held-out RNA data, totalVI and scVI were largely comparable and outperformed FA (Extended Data Fig. 2a, b). totalVI and scVI also had a comparable held-out predictive log-likelihood for the RNA data (Extended Data Fig. 2c). Finally, totalVI's performance was also stable across multiple initializations (Extended Data Fig. 2d, e).

To assess the scalability of totalVI, we concatenated all of our spleen and lymph node data (SLN-all) and recorded the training time for different sizes of subsets of this data. totalVI and scVI had similar dependence between run time and input size (Extended Data Fig. 2e). Furthermore, we observed that totalVI can readily handle large data sets, for instance, processing the complete set of approximately 33,000 cells with over 4,100 features (genes and proteins) in under one hour.

totalVI identifies and corrects for protein background

To analyze protein data in an accurate and quantitative manner, it is necessary to distinguish between true biological signal and technical bias in the protein measurement. Background is a type of technical bias that is characteristic of antibody-based measurements [6, 7, 27]. In CITE-seq data, protein measurements include non-negligible background that arises experimentally from a combination of ambient antibodies, which can be detected in empty droplets, and non-specific antibody binding, which can be detected above ambient levels in cells with no expected expression of a protein, such as CD19 in T cells (Methods, Extended Data Fig. 3a-c, g). Recent methods have described background from ambient RNA [28-30], but the presence of background is more pronounced in protein measurements (Extended Data Fig. 3d-f, Supplementary Note 3).

Previous studies of CITE-seq data derived a single decision rule for every protein, specifying the minimum number of counts required to be considered foreground by using either spiked-in negative control cells [6] or a Gaussian mixture model (GMM) to distinguish a background and foreground component for each protein [31]. Using the same boundary for all cells, however, relies on the assumption that all cells are subject to a similar background distribution of the protein in question and, in the case of a two-component GMM, that the foreground component is comparable across cell types.

totalVI instead models protein background as cell- and protein-specific. To do this, totalVI models each protein measurement as a mixture of foreground and background components that depends on the cell's representation in the latent space, and therefore the full transcriptomic and proteomic profile of that cell. The mixture is weighted by the probability that the counts of a protein in a given cell came from the background component (Fig. 1b, Methods).

To evaluate totalVI's ability to quantitatively identify protein background, we tested how well major cell types could be predicted by the foreground probability (one minus the background probability) of common marker proteins in the SLN111-D1 dataset (Methods). As a baseline for comparison, we used the assignment probabilities from a two-component GMM. For nine out of eleven known marker proteins, both totalVI and the GMM performed well at classifying cell types by marker foreground probability (ROC AUC > 0.97; Supplementary Table 2). For these proteins, such as the B cell marker CD19, the distributions of foreground and background expression were easily separated (Extended Data Fig. 3a and Supplementary Fig. 1a-d). However, for the B cell marker CD20 and the T cell marker CD28, distributions of foreground and background expression were highly overlapping (Extended Data Fig. 3b, c), and totalVI noticeably outperformed the GMM (Extended Data Fig. 3h). totalVI also performed better at distinguishing foreground and background for this set of proteins in the SLN208-D1 dataset, even after normalizing the raw data using isotype control antibodies [32] prior to fitting the GMM (Methods, Supplementary Table 3). Across all proteins, the totalVI foreground probability tended to fall near zero or one, indicating the model's certainty about most measurements (Supplementary Fig. 1e).

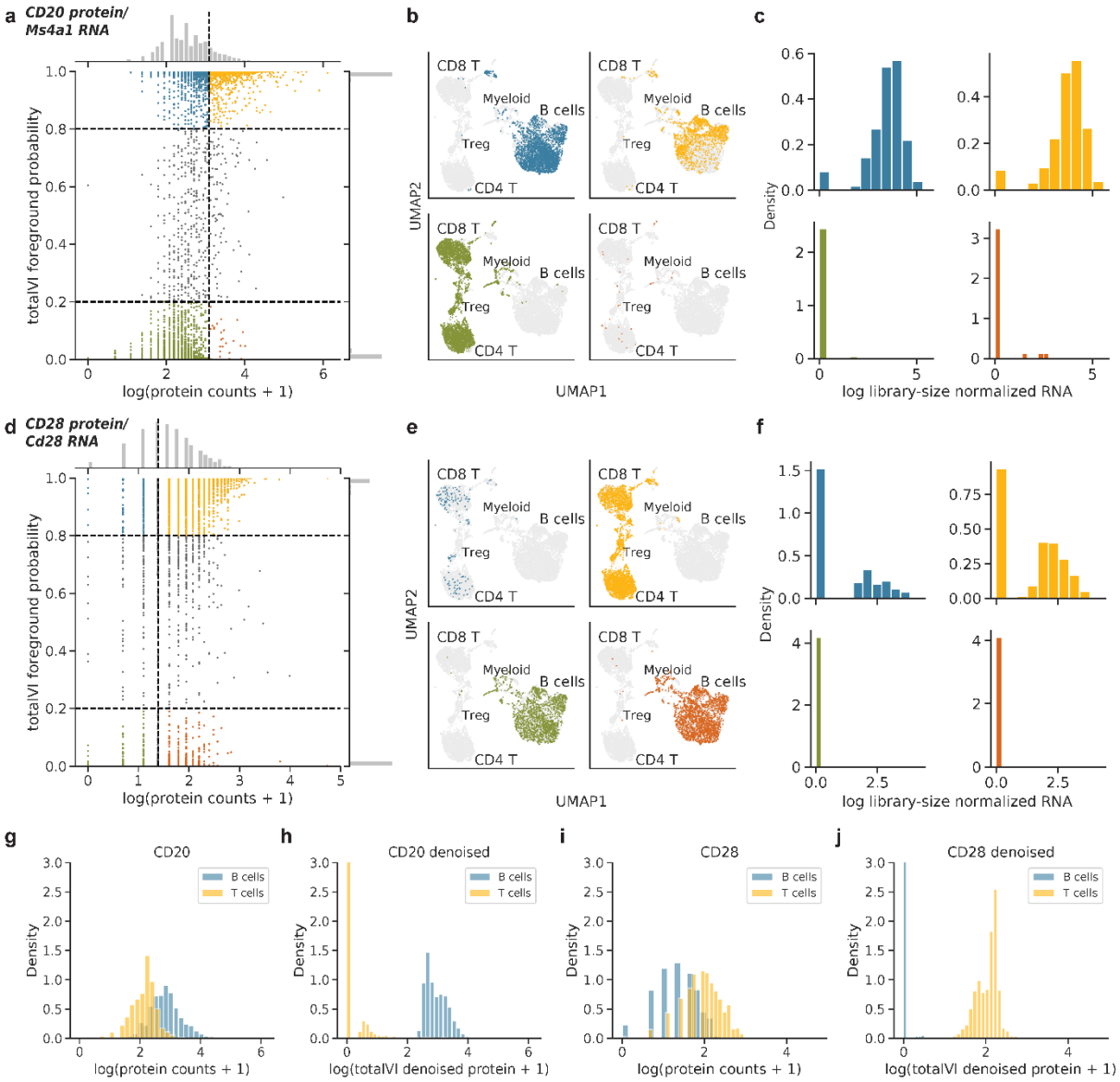


Figure 2: totalVI identifies and corrects for protein background. totalVI was applied to the SLN111-D1 dataset. **a-c**, CD20 protein (encoded by Ms4a1 RNA). **(a)** totalVI foreground probability vs log($protein\ counts + 1$). Vertical line denotes protein foreground/background cutoff determined by a GMM. Horizontal lines denote totalVI foreground probability of 0.2 and 0.8. Cells with foreground probability greater than 0.8 or less than 0.2 are colored by quadrant, while the remaining cells are gray. **(b)** UMAP plots of the totalVI latent space. Each quadrant contains cells from the corresponding quadrant of (a) in color with the remaining cells in gray. **(c)** RNA expression (log library-size normalized; Methods 4.8) for cells colored in (a). **d-f**, Same as (a-c), but for CD28 protein (encoded by Cd28 RNA). **g, h**, Distributions of log($protein\ counts + 1$) **(g)** and log($totalVI\ denoised\ protein + 1$) **(h)** for CD20 protein in B cells (blue) and T cells (yellow). y-axis is truncated at 3. **i, j**, Same as (g, h), but for CD28 protein.

Using CD20 and CD28 as examples, we see how totalVI's identification of protein foreground and background is more accurate than a single decision boundary. In the case of CD20 (encoded by Ms4a1 RNA), a GMM-based cutoff resulted in numerous false negatives (blue cells in Fig. 2a-c, Methods). These cells, identified by totalVI as having high foreground probability despite low CD20 expression, clustered with B cells and expressed Ms4a1 RNA, confirming their identity as

B cells. In contrast, cells with similarly low CD20 expression but with low totalVI foreground probability (green cells) clustered with T cells and did not express Ms4a1 (Fig. 2a-c). In the case of CD28, a GMM-based cutoff resulted in numerous false positives (red cells in Fig. 2d-f), while totalVI correctly identified that these cells with high CD28 had low foreground probability, and were in fact B cells rather than T cells. totalVI is not limited to distinguishing globally bimodal distributions (e.g., CD4 in peripheral blood mononuclear cells globally follows a trimodal distribution (Methods, Extended Data Fig. 4a, b)).

For downstream analysis, totalVI uses foreground probabilities in a quantitative manner to remove protein background. Specifically, totalVI can denoise the protein data by setting the background component to zero, while also accounting for the measurement uncertainty of the foreground component (Methods, Fig. 2g-j, Extended Data Fig. 4f, g). We use the expectation of denoised values for visualization (Extended Data Fig. 4c-e).

For statistical analyses like differential expression testing, totalVI uses distributions over the denoised values as opposed to testing directly on a denoised data matrix, which could introduce bias [33]. For analyses focused on the relationships between features, we developed a novel sampling method that controls for nuisance variation while avoiding denoising-induced artifacts (Methods). We applied this method to construct denoised feature-feature correlation matrices and found that totalVI preserved the independence of negative control genes (Extended Data Fig. 5a, b, d, e), lending confidence that downstream analysis with totalVI is not subject to spurious feature relationships arising from data denoising. Observing the correlations between proteins and their encoding RNA, we found that totalVI correlations were generally higher in magnitude than raw correlations (Extended Data Fig. 5c, f).

totalVI integrates CITE-seq datasets

We next evaluated totalVI's ability to integrate data from CITE-seq experiments that measured different sets of proteins. Integration is built into totalVI via an assumption of independence between the latent space and the batch. Consequently, totalVI produces both an integrated latent space, as well as corrected expression values. In the case of unmatched protein panels, totalVI can impute missing proteins for a particular dataset by using the information learned from those proteins in the datasets in which they were observed (Methods). We applied totalVI to the SLN111-D1 and SLN208-D2 datasets, which had a clear batch effect that was revealed by principal component analysis (Fig. 3a). We benchmarked totalVI against three state-of-the-art integration methods: Seurat v3 [34], Scanorama [35], and Harmony [36]. We assessed totalVI in the case of matched panels (using only the 111 overlapping proteins between the two panels; denoted as totalVI-intersect) and unmatched panels (using the union of the two protein panels, which results in missing data for some proteins; denoted as totalVI-union). Despite being designed for scRNA-seq, the other methods could be extended to handle CITE-seq data, though only in the case of matched panels (Methods).

We used four metrics to quantify how well each method mixed datasets along with how well they maintained the original structure of each dataset (Methods). The first two metrics (the latent mixing metric and the measurement mixing metric) quantify how well cells mix across datasets in the low-dimensional latent space and the observed expression space (per feature), respectively. The second two metrics (the feature retention metric and clustering metric) summarize how well each method

preserves each dataset's original structure, either at the feature-level through autocorrelation (feature retention metric), or at the cell-level through clusters (clustering metric). Finally, we benchmarked totalVI's accuracy of predicting protein expression in cases where measurements are available in only one of the datasets.

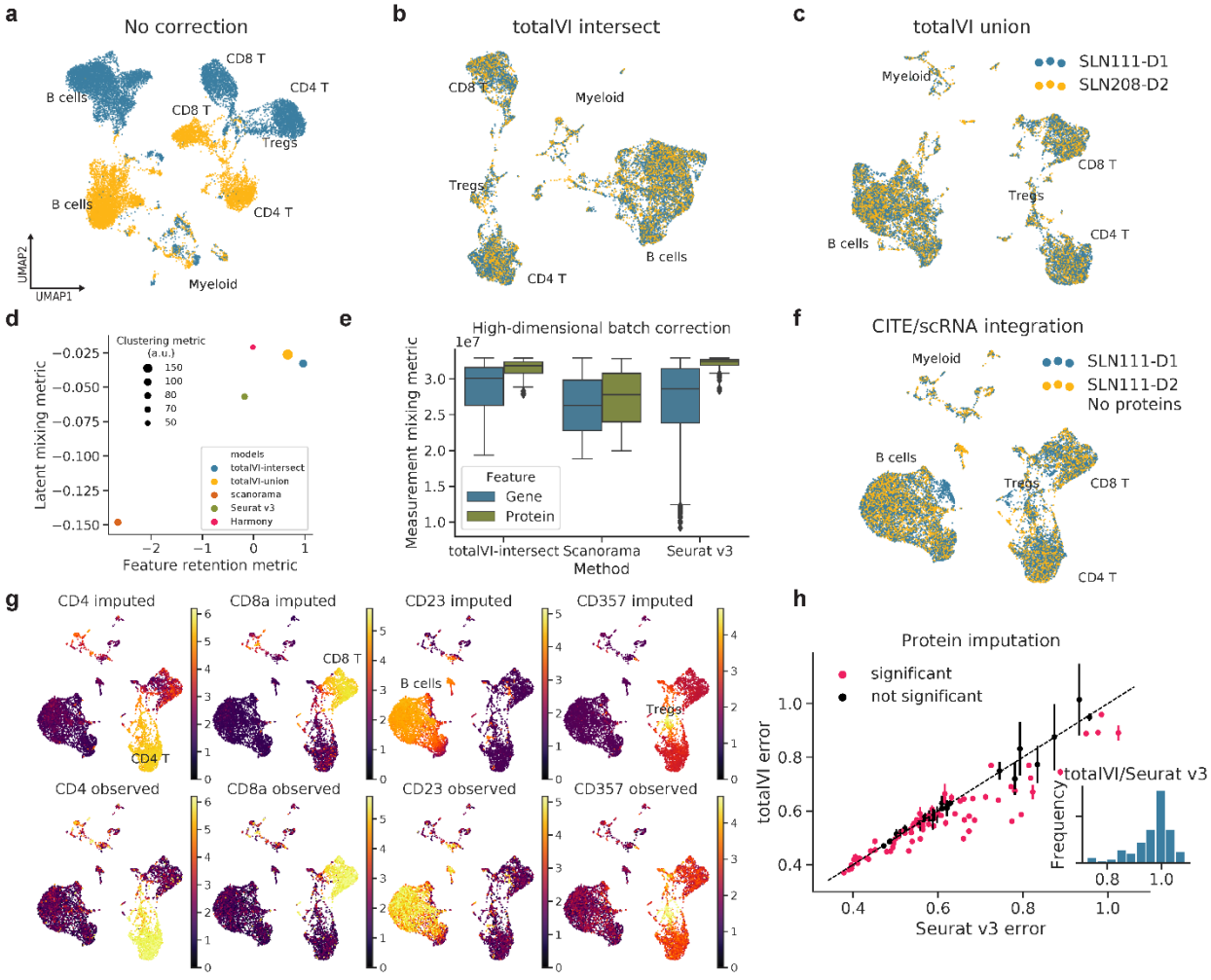


Figure 3: Benchmarking of integration methods for CITE-seq data. a-c, UMAP plots of SLN111-D1 and SLN208-D2 with no integration (PCA of paired data with intersection of protein panels), and after integration with totalVI-intersect, in which the protein panels were intersected, and totalVI-union, in which the unequal protein panels were preserved, colored by dataset. **d, e**, Performance of integration methods based on four metrics: **(d)** latent mixing metric, feature retention metric, clustering metric (displayed as point size), and **(e)** measurement mixing metric (computed for $n = 4000$ genes and $n = 111$ proteins; higher values are better for each; Methods). Box plots indicate the median (center lines), interquartile range (hinges), whiskers at 1.5x interquartile range. **f**, UMAP plot of SLN111-D1 integrated with SLN111-D2 (proteins held out) by totalVI. **g**, UMAP plots colored by totalVI imputed and observed protein expression (log scale) of key cell type markers (range 0-99th percentile of held-out values for each protein). **h**, Root mean squared error (RMSLE) of imputed versus observed protein expression (log scale) for totalVI-union and Seurat v3. totalVI performance per protein is presented as mean RMSLE with error bars representing 95% confidence intervals of the mean estimate ($n = 30$ model initializations). Proteins colored in black are not significantly different in performance, while those in red are significantly different (two-sided Student's t-test, BH-adjusted p-value < 0.05). Inset displays ratio in performance across proteins for totalVI and Seurat v3.

We found that after integration, cells of similar types were co-located in the latent space, as evidenced by the shared expression of key marker proteins like CD4, CD8a, and CD19 (Fig. 3b, c; Supplementary Fig. 2). Moreover, totalVI outperformed the other methods in the feature retention and clustering metrics, while comparing favorably in the remaining metrics (Fig. 3d, e). totalVI-union and totalVI-intersect performed similarly, indicating that the presence of missing data did not diminish totalVI's integration capabilities. We repeated this analysis on two public datasets of PBMCs (PBMC10k [23], PBMC5k [37]), which also had very different sequencing depths, and observed similarly favorable performance for totalVI (Supplementary Fig. 3a-f).

Since totalVI-union can integrate CITE-seq datasets with different protein panels, we reasoned it could also integrate a CITE-seq dataset with a standard scRNA-seq dataset that has not measured proteins and impute the missing protein measurements. We assessed this by integrating SLN111-D1 and SLN111-D2, where we held out the proteins of SLN111-D2. We first observed that totalVI can learn a biologically meaningful integrated latent representation despite the large amount of missing data (Fig. 3f). Indeed, the location of observed protein expression in the latent space revealed the same broad immune cell types. Next, we imputed the protein expression for the cells in SLN111-D2 (Methods). For key cell type marker proteins, totalVI-imputed proteins shared similar patterns of expression as the held-out observed proteins (Fig. 3g).

To further quantify imputation accuracy, we ran totalVI 30 times with resampled training sets and, for each run, computed the root mean squared log error between imputed and observed protein values. We compared totalVI to Seurat v3, which imputes protein values based on smoothing of protein values from mutual nearest RNA neighbors. The accuracy of 80 proteins was significantly different between totalVI and Seurat v3 (Student's T-test, Benjamini–Hochberg (BH)-adjusted p-value <0.05). The mean error of totalVI was better than the Seurat v3 error for approximately 68% of the 80 proteins (Fig. 3h). We also performed this task on PBMCs (Supplementary Fig. 3h, i), in which we also compared to another protein imputation method, cTP-net [38]. We found that totalVI and Seurat v3 performed more similarly, while outperforming cTP-net. For further discussion on the merits and limitations of imputing missing proteins, see Supplementary Note 4.

totalVI identifies differentially expressed genes and proteins

totalVI can leverage its estimates of uncertainty from a single model fit to detect differentially expressed features between two sets of cells while controlling for noise and other modeled technical biases like sequencing depth (RNA), background (protein), and batch effects (both). To do so, totalVI estimates a distribution over the log fold change (LFC) of expression between the two sets of cells, which is then used to quantify how well the data support a hypothesis of differential expression (using Bayes factors [15, 39, 40]; Methods).

To evaluate totalVI as a framework for differential expression (DE) analysis in the common scenario of multiple experiments, we integrated all four spleen and lymph node datasets (SLN-all; totalVI-intersect). totalVI provided a descriptive representation of this data, as inspection of established cell type markers associated clusters of cells in the latent space with immune cell types or states (Fig. 4a, Extended Data Fig. 6, Methods).

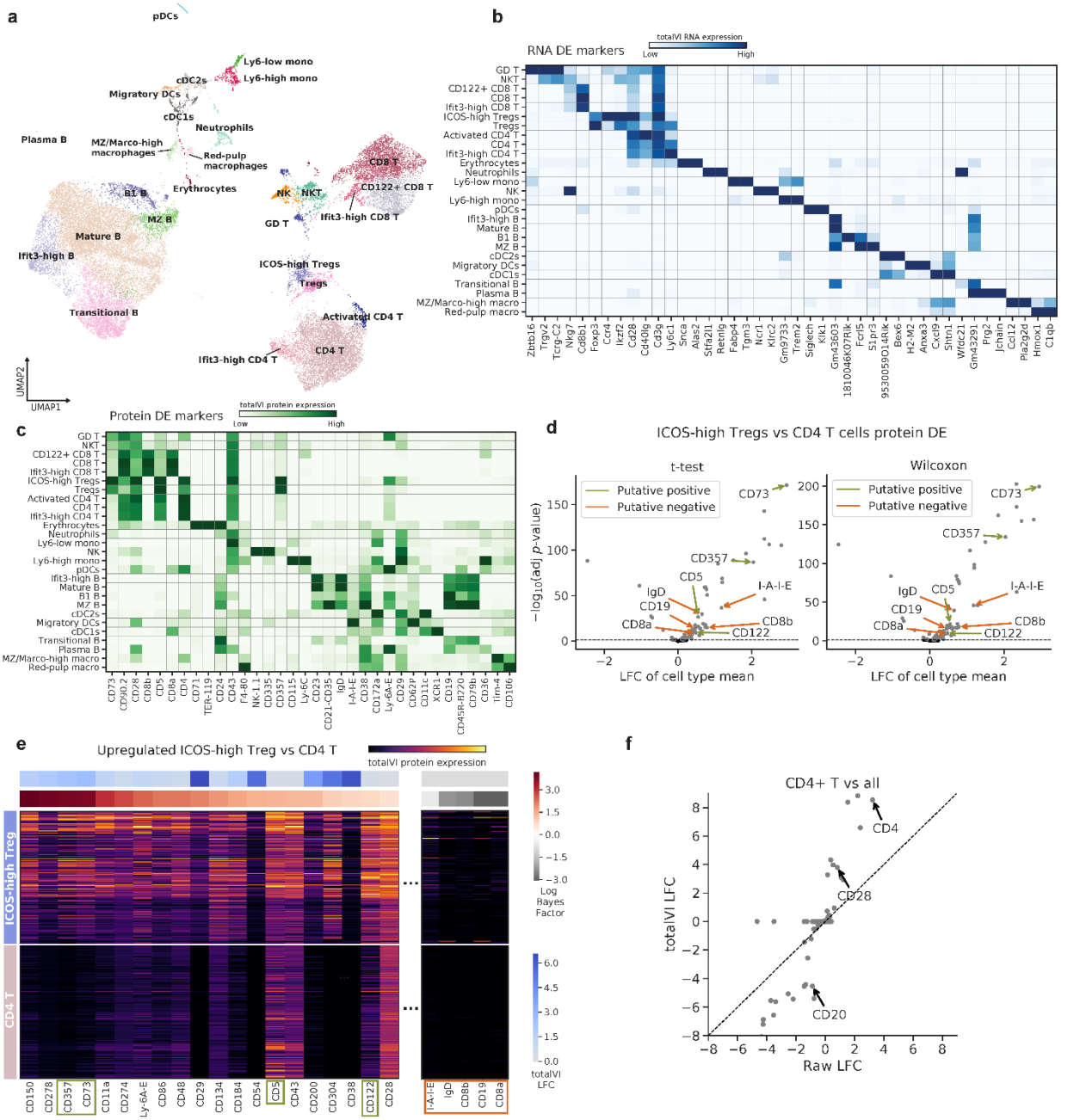


Figure 4: totalVI identifies differentially expressed genes and proteins. totalVI intersect was applied to the SLN-all dataset. **a**, UMAP plot of SLN-all, after clustering and annotating the data (Methods 4.11). **b, c**, Heatmap of markers derived from one-vs-all tests for **(b)** RNA and **(c)** proteins. For each cell type, we display the top three protein markers and top two RNA markers in terms of LFC. **d**, Volcano plot of protein differential expression test between ICOS-high Tregs and CD4 T cells for a Welch's t-test and Wilcoxon rank-sum test. Putative positives and negatives are denoted by green and orange arrows, respectively. Significant proteins (BH-adjusted p-value < 0.05) are colored in grey, all others are in black. **e**, totalVI protein expression for proteins (columns) upregulated in ICOS-high Tregs versus CD4 T cells. Cells (rows) are ordered by cluster, and subsampled to be equal in number per cluster. Columns are normalized in the range [0, 1]. The left section in the heatmap contains all the proteins called differentially expressed by totalVI with a positive log fold change. Proteins are sorted by Bayes factor (significance). The rightmost section contains the putative negatives, which are not called differentially expressed by totalVI. **f**, Comparison of log fold changes estimated by totalVI and observed in the raw data from a one-vs-all test of CD4 T cells.

Beyond markers used for annotation, we found that a totalVI one-vs-all DE test (in which one cell type is compared to all others) identified many additional features as differentially expressed (Methods, Fig. 4b, c; Supplementary Data). For example, totalVI identified the gene Klrc2 as differentially expressed in both natural killer (NK) cells and gamma/delta T cells, which has previously been shown to be upregulated in these populations relative to alpha/beta T cells [41]. For proteins, totalVI identified CD335 (NKp46) as among the top markers for NK cells, which is a canonical marker used for sorting [42], and CD43, which is associated with the development of NK cells [43].

Overall, the Bayes factors inferred by totalVI for the RNA data were highly correlated with those produced by scVI (Extended Data Fig. 7a), which has been independently evaluated [40]; therefore, we focused on evaluating the protein DE test. Throughout, we compared totalVI to two baseline methods: a Welch's t-test and a Wilcoxon rank-sum test. We also compared to a version of totalVI in which the protein background was not corrected (totalVI-wBG).

We first evaluated the extent of false positives using isotype control antibodies. As isotype controls lack target specificity, differences in their abundance between cell types likely stem from background or other technical sources of variation. Applying each method to the SLN208-D1 dataset, which contained nine isotype controls, we found that totalVI called the fewest (and often zero) isotype controls as differentially expressed in one-vs-all tests (Extended Data Fig. 7b). We next tested the reproducibility of the methods across biological replicates, finding that totalVI outperformed the baseline methods (Extended Data Fig. 7c-e). The totalVI DE test was also reproducible across experimental designs: one in which the two CITE-seq datasets had the same protein panel, and another in which proteins were measured in only one of the datasets (Extended Data Fig. 7f).

To gain further insight into the extent of false positive and false negative DE calls, we compared ICOS-high regulatory T cells (ICOS-high Tregs) and conventional CD4 T cells from SLN-all. This test is challenging because these two cell types share many of the same upregulated and downregulated features when compared with other immune cell types. Our analysis was based on a list of putative positive and negative surface proteins curated from previous studies that used flow cytometry (Methods).

We found that totalVI and the baseline methods identified these putative positives as significantly upregulated; however, the two baseline methods also incorrectly called all putative negatives as upregulated (Fig. 4d). Globally, the two baseline methods both called 78 out of 110 proteins as differentially expressed, many of which are likely the result of differences in background. While filtering proteins by the observed LFC in the baseline methods may reduce these false positives, the improvement would be limited (e.g., CD5 and IgD had similar LFCs and therefore could not be distinguished; Fig. 4d). The totalVI test, in contrast, correctly classified all putative negatives and positives (Fig. 4e), calling 28 proteins differentially expressed in total. To further support the utility of correcting for protein background, we performed this test using totalVI-wBG, which improved upon the baseline methods, but also falsely called some putative negatives as positives (Supplementary Fig. 4a).

Finally, totalVI's LFC estimates (defined as the median of the LFC distribution) better captured the underlying biological signal. For example, in a test of CD4 T cells vs all from SLN-all, the canonical marker CD4 had a higher LFC than in the raw data (Fig. 4f). Additional markers like CD28 (T cell marker) and CD20 (B cell marker), which we previously highlighted as having highly overlapping foreground and background components, had respectively higher and lower LFCs compared to LFCs derived from the raw data.

totalVI provides an interpretable latent space

Deep-learning-based methods for dimensionality reduction tend to rely on "black-box" models, making it difficult to interpret the coordinates of their inferred low-dimensional latent spaces. Despite the non-linear relationship between the totalVI latent space and the expression space, totalVI provides a way to relate each latent dimension to the expression of individual features via archetypal analysis [21, 44, 45] (Methods). Archetypes, which correspond to dimensions of the latent space, represent a summary of expression programs, the combination of which characterizes a cell. To demonstrate archetypal analysis, we ranked the features most associated with each archetype in the SLN-all dataset (Extended Data Fig. 8a, b), finding that some archetypes corresponded to specific cell types, and others captured more global variation (Extended Data Fig. 9a). For example, archetype 16 was associated with high protein expression of CD93 and CD24, which mark the transitional B cell subset (Extended Data Fig. 9b). In contrast, archetype 7 was associated with interferon-response genes such as Ifit3 and Isg20 and reflected within-cell-type variability in several subsets, including CD4 and CD8 T cells, B cells, Ly6-high monocytes, and neutrophils (Extended Data Fig. 9c and Supplementary Fig. 5). We also used archetypal analysis to evaluate the influence of proteins on the latent space, and found that all but one archetype had proteins overrepresented in its top features (Extended Data Fig. 8c). This suggests that the inclusion of proteins significantly influences representations in the totalVI latent space.

Characterization of B cell heterogeneity in the spleen and lymph nodes with RNA and proteins

We next demonstrate how a joint representation of RNA and protein can be used to characterize cell identities within a specific immune compartment and in the context of multiple samples. Here, we used the totalVI-intersect model fit on the SLN-all dataset and focused on the B cell population (Methods, Fig. 4a).

We started with characterizing cell identities using prior biological knowledge by visualizing the expression of six surface proteins commonly used for isolating B cell subsets (Fig. 5b, Supplementary Table 4). These subsets included transitional (marked by CD93 and CD24), mature (marked by IgD and CD23), B1 (marked by CD43) and marginal zone (MZ, marked by CD21) B cells. These markers stratified the B cells into groups that were largely consistent with unsupervised clustering (Methods). RNA expression of these markers followed similar patterns to the proteins they encode (Fig. 5c).

The difference in subset composition between the spleen and lymph nodes (Fig. 5d) was consistent with previous studies (Fig. 5e, [46, 47]). In particular, clusters spanned the developmental range from recent bone-marrow emigrants in the splenic transitional B cell subset to mature cells present in both tissues. As expected, the B1 and MZ B cell subsets were found primarily in the spleen.

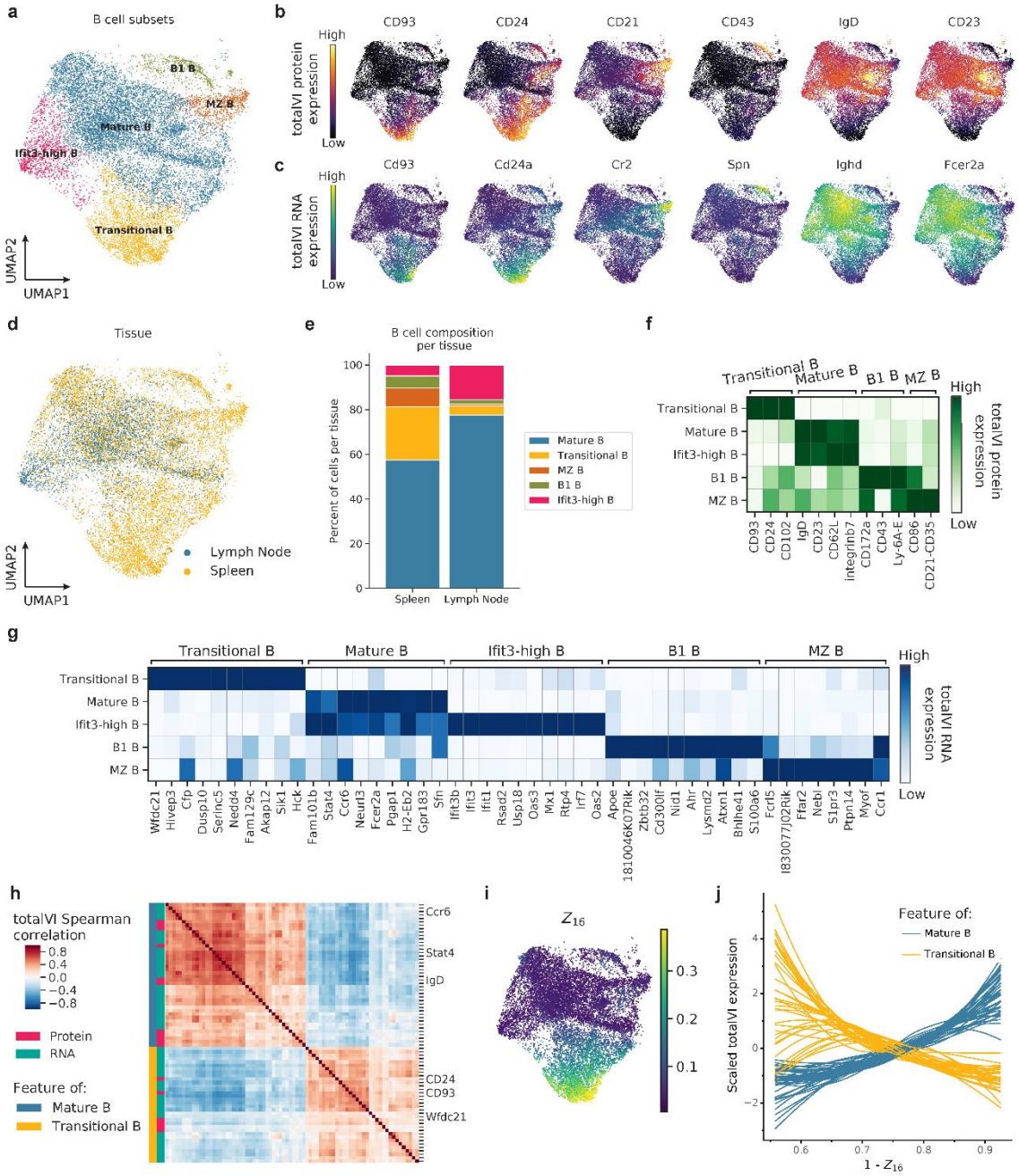


Figure 5: Characterization of B cell heterogeneity in the spleen and lymph nodes with RNA and protein. totalVI-intersect was applied to the SLN-all dataset. Data were filtered to include B cells. **a**, UMAP plot of totalVI latent space labeled by cell type. **b, c**, UMAP plots of totalVI latent space colored by **(b)** totalVI protein expression of six marker proteins and **(c)** totalVI RNA expression of the six genes that encode the corresponding proteins in (b). **d**, UMAP plot of totalVI latent space labeled by tissue. **e**, Cell type composition per tissue. **f, g**, totalVI one-vs-all differential expression test on B cell subsets filtered for significance (Methods) and sorted by the totalVI median LFC. **(f)** The top three differentially expressed proteins per subset and **(g)** the top ten differentially expressed genes per subset, arranged by the subset in which the LFC is highest. **h**, totalVI Spearman correlations in transitional B cells between RNA and proteins, which were selected as described in Methods. Features were hierarchically clustered and are labeled as either RNA or protein, and by the cell type with which the feature is associated. **i**, UMAP plot of totalVI latent space colored by Z_{16} (the totalVI latent dimension associated with transitional B cells). **j**, totalVI expression of features in (h) as a function of $(1 - Z_{16})$. Each feature was standard scaled and smoothed with a loess curve.

In a more unbiased approach, we quantified the differences between the B cell clusters with the totalVI one-vs-all DE test (Fig. 5f, g, Methods). As expected, the six known surface markers were among the top differentially expressed protein markers (Fig. 5f). Most RNA molecules encoding the marker proteins were also differentially expressed along with informative genes whose products are not present on the cell surface, such as the transcription factor Bhlhe41 that marks B1 B cells (Fig. 5g, [48]).

Globally, protein data combined with a transcriptome-wide view enabled a more refined characterization of variation within the four major sub-populations identified above by surface markers. For example, a sub-population of mature B cells labeled here as Ifit3-high B cells expressed all of the protein and RNA markers of mature B cells and could not be clearly distinguished from the remaining mature B cells based on protein data alone (maximum LFC across all proteins was less than 0.19). Nevertheless, with transcriptome-wide DE analysis, this cluster could be distinguished as a sub-type of mature cells by the elevated expression of interferon response genes (Fig. 5g). This observation was supported by a gene signature analysis with Vision [49], which identified two interferon response signatures enriched in the Ifit3-high B cell cluster (Methods, Supplementary Fig. 5a, b). The expression of interferon response genes was not expected since no inflammation was induced, however we found the Ifit3-high B cell cluster as well as Ifit3-high T cell clusters to be represented in both biological replicates, and therefore took it to capture part of the biology in the SLN-all dataset (Supplementary Fig. 5c, d).

Next, we explored the variability within transitional B cells and its relationship with B cell development. Interestingly, latent dimension 16 (Z_{16}) captured a gradual transition within this cluster: from a small population of Rag1 expressing cells (indicating early development [46]) to cells that were closer to the mature cluster (Fig. 5i, Extended Data Fig. 10a, b). To explore how development from transitional to mature B cells may be associated with coordinated changes in gene and protein expression, we calculated the totalVI Spearman correlations separately within transitional and mature B cells for a set of features that distinguished the two subsets (Methods). Hierarchical clustering of the correlation matrix within the transitional B cells clearly stratified these features into two anti-correlated modules: one associated with transitional B cells and the other with mature B cells (Fig. 5h). These modules, however, were not present in mature B cells, indicating that the apparent coordination may be a characteristic of the transitional state (Extended Data Fig. 10c). Within transitional B cells, we found that the features in the two modules significantly correlated with the axis of maturation captured by Z_{16} (Extended Data Fig. 10d). Along this axis, features in the transitional module decreased while those in the mature module increased (Fig. 5j, Methods). These results point to a program of transitional B cell maturation that consists of coordinated activation and repression of multiple genes and proteins, leading to a gradual transition in cell state that is captured by a specific dimension of the totalVI latent space.

Discussion

totalVI is a scalable, probabilistic framework for end-to-end analysis of paired transcriptome and protein measurements in single cells. Like other multi-omics analysis methods [31, 50, 51], totalVI assumes that RNA and protein measurements are generated from the same latent space of cells that captures their state. A distinction of totalVI is that it explicitly models modality-specific technical factors like protein background, which we demonstrated can enable a denoised view of

the data and more accurate differential expression results. totalVI is also unique in its ability to handle missing protein data, which enables integration with growing public data resources like the Human Cell Atlas [8].

Beyond the characterization of cell types, totalVI can also uncover relationships between RNA and protein molecules within a cell. For example, totalVI could be used to investigate the relationship between the level of an RNA transcript and the level of its encoded protein in different biological settings, which remains an open question [52]. We found that the totalVI correlations were higher in magnitude than raw correlations across the majority of RNA-protein pairs, suggesting that the low correlations observed previously [6, 7] could have been due to technical noise. Future work quantifying correlations and regulatory relationships between RNA and protein features could inform our understanding of signal transduction pathways or transcription and translation dynamics [53].

While the totalVI model was designed to reflect our understanding of the CITE-seq experimental data-generating process (Supplementary Note 3), totalVI can also be used to inform experimental design. For instance, totalVI could help identify antibody titrations or experimental methods that improve signal-to-noise. totalVI could also identify sequencing depths for RNA and protein libraries that balance the information gained per measurement in various analysis tasks with the cost of additional sequencing [54, 55].

Through a single pipeline that jointly analyzes paired RNA and protein measurements, totalVI simplifies data analysis and interpretation that would otherwise be conducted in separate pipelines whose disparate results must be reconciled post hoc. totalVI is available through the scvi-tools software package, which connects it with the popular Scanpy [56] and Seurat [34] pipelines, and enables analysis on free cloud computing environments like Google Colab. The flexibility and scalability of totalVI make it easily applicable to future datasets with larger protein panels, and enable extensions that incorporate additional paired measurements. For example, we expect totalVI to naturally handle intracellular proteins measured with barcoded antibodies. Further additions of modalities like chromatin accessibility [57] or clonotype features [58] can also be implemented within the totalVI codebase with consideration of the modality-specific likelihood. By combining multiple views of cellular processes, totalVI can reveal a more complete picture that redefines cell states and elucidates mechanistic relationships between molecular components of the cell.

Acknowledgements
We thank Ellen Robey, Lydia Lutes, and Derek Bangs for help designing experiments. We thank BioLegend Inc. and their proteogenomics team, especially Bertrand Yeung, Andre Fernandes, Qing Gao, Hong Zhang, Tse Shun Huang, for providing reagents and expertise and for help with sample preparation, library generation, and sequencing of CITE-seq libraries. We thank David DeTomaso for general data analysis advice, and Pierre Boyeau, Achille Nazaret, and Galen Xing for help with integrating totalVI in the scvi-tools package. We thank members of the Streets and Yosef laboratories for helpful feedback. Research reported in this manuscript was supported by the NIGMS of the National Institutes of Health under award number R35GM124916 (A.S), the Chan-Zuckerberg Foundation Network under grant number 2019-02452 (N.Y.), and the National Institutes of Mental Health under grant number U19MH114821 (N.Y.). A.G. is supported by NIH

Training Grant 5T32HG000047-19. Z.S. is supported by the National Science Foundation Graduate Research Fellowship. N.Y. was supported by the Koret-Berkeley-Tel Aviv (KBT) Initiative in Computational Biology. A.S. and N.Y. are Chan Zuckerberg Biohub investigators.

Author contributions

A.G. and Z.S. contributed equally. A.G., Z.S., A.S., and N.Y. designed the study. A.G., Z.S, R.L., J.R., and N.Y. conceived of the statistical model. A.G. implemented the totalVI software with input from R.L. K.L.N. designed and produced antibody panels and provided input on the study. Z.S. designed and led experiments with input from A.S. and N.Y. A.G. and Z.S. designed and implemented analysis methods and applied the software to analyze the data with input from A.S. and N.Y. A.S. and N.Y. supervised the work. A.G., Z.S., R.L., J.R., A.S., and N.Y. participated in writing the manuscript.

Ethics declaration

K.L.N. is an employee of BioLegend Inc. The other authors declare no competing interests.
References

References

1. Stubbington, M. J. T., Rozenblatt-Rosen, O., Regev, A. & Teichmann, S. A. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**, 58–63 (2017).
2. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* (2017) doi:10.1038/nri.2017.76.
3. Labib, M. & Kelley, S. O. Single-cell analysis targeting the proteome. *Nature Reviews Chemistry* **4**, 143–158 (2020).
4. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* (2016) doi:10.1038/nbt.3711.
5. Efremova, M. & Tiechmann, S. A. Computational methods for single-cell omics across modalities. *Nature Methods* (2020).
6. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* (2017) doi:10.1038/nmeth.4380.
7. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology* (2017) doi:10.1038/nbt.3973.
8. Regev, A. *et al.* The Human Cell Atlas. *eLife* (2017) doi:10.7554/eLife.27041.
9. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* (2017) doi:10.1038/nature21350.
10. Todorovic, V. Single-cell RNA-seq—now with protein. *Nature Methods* **14**, 1028–1029 (2017).
11. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* **9**, 1–12 (2017).
12. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology* **37**, 1458–1465 (2019).
13. Praktiknjo, S. D. *et al.* Tracing tumorigenesis in a solid tumor model at single-cell resolution. *Nature Communications* **11**, 991 (2020).

14. Kotliarov, Y. *et al.* Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine* **26**, 618–629 (2020).
15. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018).
16. Levitin, H. M. *et al.* De novo gene signature identification from single-cell RNA -seq with hierarchical Poisson factorization. *Molecular Systems Biology* **15**, (2019).
17. Azizi, E., Prabhakaran, S., Carr, A. & Pe'er, D. Bayesian inference for single-cell clustering and imputing. *Genomics and Computational Biology* (2017) doi:10.18547/gcb.2017.vol3.iss1.e46.
18. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* (2018) doi:10.1038/s41467-017-02554-5.
19. Blei, D. M. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application* (2014) doi:10.1146/annurev-statistics-022513-115657.
20. Kingma, D. P. & Welling, M. Auto-Encoding variational Bayes. in *International conference on learning representations* (2014).
21. Cutler, A. & Breiman, L. Archetypal analysis. *Technometrics* (1994) doi:10.1080/00401706.1994.10485840.
22. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology* (2018) doi:10.1186/s13059-018-1603-1.
23. 10X Genomics. 10k PBMCs from a Healthy Donor - gene expression and cell surface protein. (2018).
24. 10X Genomics. 10k Cells from a MALT Tumor - gene expression and cell surface protein. (2018).
25. Gelman, A., Meng, X. L. & Stern, H. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* (1996).
26. Kuleshov, V., Fenner, N. & Ermon, S. Accurate uncertainties for deep learning using calibrated regression. in *International conference on machine learning* (2018).
27. Hulspas, R., O'Gorman, M. R. G., Wood, B. L., Gratama, J. W. & Sutherland, D. R. Considerations for the control of background fluorescence in clinical flow cytometry. *Cytometry Part B: Clinical Cytometry* (2009) doi:10.1002/cyto.b.20485.
28. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology* **21**, 57 (2020).
29. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv* 303727 (2018) doi:10.1101/303727.
30. Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv* (2019) doi:10.1101/791699.
31. Ngo Trong, T. *et al.* Semisupervised generative autoencoder for single-cell data. *Journal of Computational Biology* (2019) doi:10.1089/cmb.2019.0337.
32. Li, B. *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nature Methods* **17**, 793–798 (2020).

33. Andrews, T. S. & Hemberg, M. False signals induced by single-cell imputation. *F1000Research* (2019) doi:10.12688/f1000research.16613.2.
34. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902 (2019).
35. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology* (2019) doi:10.1038/s41587-019-0113-3.
36. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, (2019).
37. 10X Genomics. 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry). (2019).
38. Zhou, Z., Ye, C., Wang, J. & Zhang, N. R. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications* **11**, 1–10 (2020).
39. Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the American Statistical Association* **90**, 773–795 (1995).
40. Boyeau, P. *et al.* Deep Generative Models for Detecting Differential Expression in Single Cells. in *Machine learning in computational biology* (2019). doi:10.1101/794289.
41. Bezman, N. A. *et al.* Molecular definition of the identity and activation of natural killer cells. *Nature Immunology* **13**, 1000–1008 (2012).
42. Walzer, T. *et al.* Identification, activation, and selective in vivo ablation of mouse NK cells via NKp46. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3384–3389 (2007).
43. Gordon, S. M. *et al.* The transcription factors T-bet and Eomes control key checkpoints of natural killer cell maturation. *Immunity* **36**, 55–67 (2012).
44. Korem, Y. *et al.* Geometry of the gene expression space of individual cells. *PLoS Computational Biology* **11**, 1–27 (2015).
45. Dijk, D. van *et al.* Finding archetypal spaces for data using neural networks. *arXiv* (2019).
46. Thomas, M. D., Srivastava, B. & Allman, D. Regulation of peripheral B cell maturation. *Cellular Immunology* **239**, 92–102 (2006).
47. Loder, F. *et al.* B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. *Journal of Experimental Medicine* **190**, 75–89 (1999).
48. Kreslavsky, T. *et al.* Essential role for the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nature Immunology* **18**, 442–455 (2017).
49. DeTomaso, D. *et al.* Functional interpretation of single cell similarity maps. *Nature Communications* (2019).
50. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics* **7**, 523–542 (2013).
51. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14**, 1–13 (2018).
52. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
53. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology* **21**, 1–6 (2020).

54. Svensson, V., Beltrame, E. da V. & Pachter, L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv* (2019) doi:10.1101/762773.
55. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems* **2**, 239–250 (2016).
56. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology* (2018) doi:10.1186/s13059-017-1382-0.
57. Clark, S. J. *et al.* ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nature Communications* **9**, 1–9 (2018).
58. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods* **16**, 409–412 (2019).

Methods

The totalVI model

totalVI estimates a conditional distribution for cell n, $p_v(x_n, y_n \mid s_n)$, in which x_n is the G-dimensional vector of observed RNA counts (G genes), y_n is the T-dimensional vector of observed protein counts (T proteins) and s_n is the B-dimensional one-hot vector describing the batch index (experiment identifier). In total, there are N cells. We use v to refer to the set of all generative parameters, which are described throughout this section. This distribution is estimated using the framework of variational autoencoders (VAE; [20]).

We begin by describing the generative process, for which a graphical summary is in Supplementary Fig. 6 and an algorithmic summary is in Algorithm 1. We then describe the inference procedure, as well as how downstream analysis tasks are directly linked to posterior queries of the model.

Priors

The latent cell representation $z_n \sim \text{LogisticNormal}(0, I)$, where the logistic normal distribution is a distribution over the probability simplex. This specification, which has also been applied in the context of linear VAEs for scRNA-seq [59], enables cells to be interpreted with archetypal analysis. Typically in VAEs, z_n follows an isotropic normal distribution, which is chosen for computational convenience [20]. In this setting, a logistic normal distribution arises as transforming a sample from a normal distribution with a softmax function. For all experiments, we set z_n to 20 dimensions. We discuss the choice of number of latent dimensions in Supplementary Note 1.

The latent RNA size factor $\ell_n \mid s_n \sim \text{LogNormal}(\ell_\mu^\top s_n, \ell_{\sigma^2}^\top s_n)$, where $\ell_\mu \in \mathbb{R}^B$ and $\ell_\sigma^2 \in \mathbb{R}_+^B$ are set to the empirical mean and variance of the log RNA library size (defined as total RNA counts of a cell) per batch. We use a protein-specific prior for the protein background intensity, where $\beta_{nt} \mid s_n \sim \text{LogNormal}(c_t^\top s_n, d_t^\top s_n)$. The parameters for the background intensity, $c_t \in \mathbb{R}^B$ and $d_t \in \mathbb{R}_+^B$, are protein specific and are treated as model parameters learned during inference. This prior is motivated by the observation that some component of the background is due to ambient antibodies. By being batch specific, these priors on ℓ_n and β_n account for differences in sequencing

depth between datasets. A prior can also be thought of as regularizing the posterior distribution, thus reducing the influence of outliers [60]. The selection of prior distribution was guided by the computational tractability and by properties that are of interest (e.g., non-negativity).

RNA likelihood

Given z_n, ℓ_n, and s_n, an observed expression level x_{ng} follows a negative binomial distribution, which we present here as a Gamma-Poisson mixture:

$$\rho_n = f_\rho(z_n, s_n) \quad (1)$$

$$w_{ng} \mid z_n, \ell_n, s_n \sim \text{Gamma}\left(\theta_g, \ell_n \rho_{ng}\right) \quad (2)$$

$$x_{ng} \mid w_{ng} \sim \text{Poisson}\left(w_{ng}\right) \quad (3)$$

The gamma distribution is parameterized by its shape and mean. The mean is equal to $\ell_n \rho_{ng}$, where ℓ_n, a scaling factor, is multiplied by ρ_{ng}, interpreted as a normalized gene frequency (because ρ_n is nonnegative and sums to one). ρ_n is the output of a neural network f_ρ, which takes z_n and s_n as input (Algorithm 1).

Integrating out w_{ng} results in the following conditional distribution:

$$x_{ng} \mid z_n, \ell_n, s_n \sim \text{NegativeBinomial}(\ell_n \rho_{ng}, \theta_g). \quad (4)$$

The parameter θ_g, which is the shape of the gamma distribution, is also the inverse dispersion of the negative binomial (Supplementary Note 5). We perform inference on the model with w_{ng} integrated out. We also treat θ_g as a model parameter learned during inference. Overall, this likelihood is equivalent to that presented in scVI [15], without zero-inflation. The negative binomial distribution has been shown to adequately handle the limited sensitivity and over-dispersion that are characteristic of this data [61].

Protein likelihood

To capture observed protein counts arising from the background or foreground, we model y_{nt} with a negative binomial mixture, given z_n, β_n and s_n. This conditional distribution is described by the following process:

$$\pi_n = h_\pi(z_n, s_n) \quad (5)$$

$$\alpha_n = g_\alpha(z_n, s_n) \quad (6)$$

$$v_{nt} \mid z_n, s_n \sim \text{Bernoulli}(\pi_{nt}) \quad (7)$$

$$r_{nt} \mid v_{nt}, \beta_{nt}, z_n, s_n \sim \text{Gamma}(\phi_t, v_{nt}\beta_{nt} + (1 - v_{nt})\beta_{nt}\alpha_{nt}) \quad (8)$$

$$y_{nt} \mid r_{nt} \sim \text{Poisson}(r_{nt}) \quad (9)$$

Here v_{nt} controls which mixture component generates the counts. Its parameter, π_{nt}, is the output of the neural network $h_\pi(z_n, s_n)$. Notably, α_{nt}, which is the output of the neural network $g_\alpha(z_n, s_n)$, is greater than one. This ensures that one of the mixture components is always larger than the other, allowing us to interpret one component as background and one component as foreground. Furthermore, π_{nt} is interpreted as the probability that any cell-protein pair has observed counts due to background alone. For one mixture component, $y_{nt} \mid z_n, \beta_{nt}, s_n, v_{nt}$

follows a negative binomial distribution, as can be seen by integrating out r_{nt}. Finally, integrating out v_{nt} too shows that y_{nt} given z_n and s_n follows a negative binomial mixture distribution, where ϕ_t is a protein-specific inverse dispersion parameter.

Algorithm 1: The totalVI generative model. The gamma distribution is parameterized by its shape and mean. Let ν be the set of model parameters described here. A dataset has G genes and T measured proteins.

Define: Neural networks
$f_\rho(z_n, s_n) : \Delta^{K-1} \times \{0,1\}^B \to \Delta^{G-1}$, (Softmax output activation)
$g_\alpha(z_n, s_n) : \Delta^{K-1} \times \{0,1\}^B \to [1, \infty)^T$, (ReLU + 1 output activation)
$h_\pi(z_n, s_n) : \Delta^{K-1} \times \{0,1\}^B \to (0,1)^T$ (Sigmoid output activation)
Require: Inverse dispersion parameters $\theta \in \mathbb{R}_+^G, \phi \in \mathbb{R}_+^T$. Neural network parameters.
for *each cell n* **do**
  $z_n \sim \text{LogisticNormal}(0, I)$ K-dim. cellular state variable
  $\rho_n = f_\rho(z_n, s_n)$ G-dim. RNA frequency
  $\alpha_n = g_\alpha(z_n, s_n)$ T-dim. foreground increment protein scaling
  $\pi_n = h_\pi(z_n, s_n)$ T-dim. mixture parameter
  $\ell_n \sim \text{Lognormal}\left(\ell_\mu^\top s_n, \ell_{\sigma^2}^\top s_n\right)$ Cell scaling factor for RNA
  for *each gene g* **do**
    $w_{ng} \sim \text{Gamma}\left(\theta_g, \ell_n \rho_{ng}\right)$
    $x_{ng} \sim \text{Poisson}\left(w_{ng}\right)$
  for *each protein t* **do**
    $\beta_{nt} \sim \text{Lognormal}(c_t^\top s_n, d_t^\top s_n)$ Scalar background mean
    $v_{nt} \sim \text{Bernoulli}(\pi_{nt})$ Scalar mixture assignment
    if $v_{nt} = 1$ **then**
      $r_{nt} \sim \text{Gamma}(\phi_t, \beta_{nt})$
      $y_{nt} \sim \text{Poisson}(r_{nt})$
    else
      $r_{nt} \sim \text{Gamma}(\phi_t, \beta_{nt}\alpha_{nt})$
      $y_{nt} \sim \text{Poisson}(r_{nt})$

Inference for totalVI

Inference in the case of fully observed proteins

The model evidence, $p_\nu(x_{1:N}, y_{1:N} \mid s_{1:N})$, cannot be computed as the integrals are analytically intractable, so Bayes rule cannot be directly applied to find the posterior distribution. Therefore, we use variational inference [62] to approximate the posterior distribution with a distribution having the following factorization:

$$q_\eta(\beta_n, z_n, \ell_n \mid x_n, y_n, s_n) := q_\eta(\beta_n \mid z_n, s_n) q_\eta(z_n \mid x_n, y_n, s_n) q_\eta(\ell_n \mid x_n, y_n, s_n). \quad (10)$$

Here η is the set of parameters of an inference network, commonly called the *encoder* – a neural network that takes a cell's combined expression as input and outputs the parameters of the approximate posterior (e.g., mean and variance). Factors of the posterior approximation share the

same family as their respective priors (e.g., $q(\beta_n \mid z_n, s_n)$ is lognormal). The approximate posterior $q_\eta(z_n \mid x_n, y_n, s_n)$, whose expectation we use as the latent cell representation, is integral to many cell-level and feature-level analyses.

For the likelihoods, as described previously, we integrate out the latent variables v_{nt}, r_{nt} and w_{ng} (Algorithm 1), yielding $p_\nu(y_{nt} \mid z_n, \beta_{nt}, s_n)$, which is a mixture of negative binomials and $p_\nu(x_{ng} \mid z_n, s_n, \ell_n)$, which is a negative binomial distribution.

The evidence lower bound (ELBO) [62] of $\log p_\nu(x_{1:N}, y_{1:N} | s_{1:N})$ is optimized with respect to the variational parameters η and model parameters ν using stochastic gradients [20]. In other words, the model parameters and approximate posterior parameters are learned simultaneously. In the VAE framework, the generative neural network is referred to as the *decoder*. Each iteration of training consists of randomly choosing a mini-batch of data (256 cells), estimating the ELBO based on this mini-batch, and updating the parameters via automatic differentiation operators. The terms corresponding to Kullback-Leibler divergences of the ELBO (Supplementary Note 6) follow a deterministic warm-up scheme [63], which helps to avoid shallow local maxima. We use the Adam optimizer [64] with weight decay to update the model parameters. Learning rate reductions and early stopping are performed based on the ELBO of a validation set. As a result of mini-batching, totalVI's memory usage is constant in the number of features in the dataset and number of neural network parameters. For example, in the runtime experiment presented in Extended Data Fig. 2f, totalVI used a constant 753 megabytes of memory on an NVIDIA Titan XP GPU. totalVI's runtime is linear in the number of cells and linear in the number of features; however, as we use early stopping, convergence may vary with the dataset size.

All neural networks are feedforward and use standard activations (e.g., exponential, softmax, sigmoid) to encode the variational and generative distributions. We use the same hyperparameters for all of our experiments. Supplementary Note 6 gives further implementation details.

Inference in the case of missing proteins

Here we adapted the training procedure from [65] to handle missing protein data. As any single batch may correspond to an experiment that used a different protein panel (or no proteins in the case of a scRNA-seq experiment), the missingness of protein features depends on the batch index s_n. Further, suppose all batches share the same set of genes. Across all batches, there are T proteins. For cell n, we denote the observed protein expressions y_n^{obs} and the unobserved protein expressions y_n^{mis}. The log likelihood of the observed data decomposes as

$$\log p_\nu(x_{1:N}, y_{1:N}^{\text{obs}}, s_{1:N}) = \sum_{n=1}^{N} log\, p_\nu(x_n, y_n^{\text{obs}} \mid s_n) \qquad (11)$$

The generative process for the observed data is the same as in Algorithm 1, with appropriate modification to only generate the features present in a particular batch. Thus, ν is the same set of model parameters described previously. Again, we use variational inference to approximate the posterior distribution with the distribution in Equation 10. In fact, all approximate posteriors share

the same encoder parameters η. We optimize the ELBO of Equation 11 similarly to the procedure used when there is no missing data (i.e., we optimize the ELBO with respect to the model parameters ν and variational parameters η). To handle mismatched dimensions in the encoder, we substitute zeros for missing proteins, and for the decoder, we only calculate the ELBO terms corresponding to observed data [66]. Therefore, this procedure naturally extends to the case when there is no observed protein data for a cell n, which would be the case when the cell is obtained from a scRNA-seq experiment. Since the quality of missing protein imputation depends on (i) the goodness of fit of totalVI to the protein for the data in which it was observed and (ii) the statistical distance of the aggregated posterior distributions of z_n for each of the batches [65, 67], we add a domain adaptation regularization term to the ELBO when training [68]. A scaling factor on this regularization term decays from one to zero early in training.

Posterior predictive distributions linked to downstream tasks

For tasks like differential expression, denoising, and finding correlations, totalVI estimates functionals of posterior predictive distributions [19]. Define $C_n = \{x_n, y_n, s_n\}$ as the set of observed data for cell n. First, consider the connection between the posterior predictive distribution of RNA data to totalVI denoised RNA expression. The posterior predictive RNA expression x_{ng}^* for gene g given C_n is distributed following:

$$p(x_{ng}^* \mid C_n) \approx \int p_\nu(x_{ng}^* \mid z_n, l_n, s_n) q_\eta(z_n, l_n \mid C_n) dz_n dl_n, \quad (12)$$

To produce denoised RNA expression, we compute the posterior predictive mean of x_{ng}^*. To further control for variation due to ℓ_n, we condition on $\ell_n = 1$. By the law of total expectation,

$$\mathbb{E}_{p(x_{ng}^* \mid C_n, \ell_n = 1)}[x_{ng}^*] = \mathbb{E}_{q_\eta(z_n \mid C_n)}\left[\mathbb{E}_{p_\nu(x_{ng}^* \mid z_n, s_n, \ell_n = 1)}[x_{ng}^*]\right] \quad (13)$$

$$= \mathbb{E}_{q_\eta(z_n \mid C_n)}[\rho_{ng}], \quad (14)$$

where ρ_{ng} is the expectation of the RNA likelihood with the additional condition that $\ell_n = 1$.

For each cell n, we can compute the denoised RNA expression by averaging samples of ρ_n generated by the following process:

1. Sample z_n from $q_\eta(z_n \mid C_n)$
2. Set $\rho_n = f_\rho(z_n, s_n)$

There are two important considerations for these posterior predictive distributions. First, we use the approximate posterior as a surrogate for the posterior. Second, these posterior predictive distributions are not tractable to compute in closed form, so we can only sample from them with ancestral sampling. Functionals of the posterior are computed using Monte Carlo integration.

Denoised protein expression

After training the model, we can generate "denoised" protein expression – protein expression effectively absent of background and controlled for sampling noise. Consider the perturbed protein generative process in which we set the background intensity to zero:

$$v_{nt} \mid z_n, s_n \quad \sim \quad \text{Bernoulli}(\pi_{nt}) \tag{15}$$

$$\tilde{r}_{nt} \mid v_{nt}, \beta_{nt}, z_n, s_n \quad \sim \quad \begin{cases} \text{Gamma}(\phi_t, \beta_{nt}\alpha_{nt}) & \text{if } v_{nt} = 0 \\ \delta_0 & \text{if } v_{nt} = 1 \end{cases} \tag{16}$$

Here δ_0 is a point mass distribution at 0. After marginalizing out v_{nt}, $\tilde{r}_{nt} \mid z_n, s_n, \beta_{nt}$ follows a zero-inflated Gamma distribution with mean $(1 - \pi_{nt})\beta_{nt}\alpha_{nt}$.

For denoising, we return the posterior predictive mean of $\tilde{r}_{nt}$. Indeed, the posterior predictive mean is equal to $(1 - \pi_{nt})\beta_{nt}\alpha_{nt}$ averaged over many posterior samples of $q(\beta_{nt}, z_n \mid C_n)$. In other words, we return the foreground mean, weighted by the probability that the observation was derived from the foreground. This can also be stated as subtracting the expected background from the expected total signal.

Missing protein imputation

To impute protein expression y_{nt}^* for cell n and protein t missing in batch s_n, but that is observed in a batch $s' \neq s_n$, do the following:

1. Sample z_n from $q_\eta(z_n \mid C_n)$
2. Sample β_{nt} from $q_\eta(\beta_{nt} \mid z_n, s = s')$
3. Sample y_{nt}^* from $p_v(y_{nt}^* \mid z_n, \beta_n, s = s')$

This process returns samples of $p\left(y_{nt}^* \mid C_n, s = s'\right)$. Intuitively, we encode the cell into the latent space, which is designed to mix the batches (i.e., be an integrated low-dimensional representation of the data), and obtain the parameters for the protein likelihood (decode) conditioned on the cell being in batch $s = s'$. Thus, the quality of imputation relies on how well batches mix in the totalVI latent space. Ultimately, we report the expected value of the imputed distribution

$$\mathbb{E}_{p(y_{nt}^* \mid C_n, s = 1)}[y_{nt}^*] = \mathbb{E}_{q_\eta(z_n \mid C_n)}\left[\mathbb{E}_{p(y_{nt}^* \mid z_n, s = 1, \beta_{nt})}[y_{nt}^*]\right] \tag{17}$$

We may also impute the denoised expression, by exchanging $p_v(y_{nt}^* \mid z_n, \beta_n, s)$ with $p_v(\tilde{r}_{nt} \mid z_n, \beta_n, s)$. This change would additionally remove the protein background contribution to the prediction.

Differential expression

With a single model fit, totalVI can detect differentially expressed features between sets of cells, i.e., the model does not need to be retrained for every test. Here we use the Bayesian framework of [40] to detect differential expression (DE) of genes and proteins. Let

$$\lambda_{a,b} := \Lambda(z_a, z_b, s_a, s_b) := \log_2 \rho_a - \log_2 \rho_b \tag{18}$$

be the log fold change (LFC) of RNA expression between cells a and b. Then the probability that gene g is differentially expressed (DE) is

$$p\left(\left|\lambda_{a,b}^g\right| \geq \delta \mid C_a, C_b\right) \approx \int \mathbb{1}\{\left|\lambda_{a,b}^g\right| \geq \delta\} q(z_a \mid C_a) q(z_b \mid C_b) dz_a dz_b, \tag{19}$$

where δ is a threshold for the effect size. Intuitively, we are measuring the fraction of posterior samples that the absolute LFC greater than or equal to δ. For all experiments we set $\delta = 0.2$. We compare the DE probability to the probability that the LFC is in the null region $\left|\lambda_{a,b}^{g}\right| < \delta$ using a Bayes factor:

$$\mathrm{BF}_{a,b}^{g} = \frac{p\left(\left|\lambda_{a,b}^{g}\right| \geq \delta \mid C_a, C_b\right)}{p\left(\left|\lambda_{a,b}^{g}\right| < \delta \mid C_a, C_b\right)}. \qquad (20)$$

This can also be extended to groups of cells. Let $A = a_1, a_2, \ldots, a_m$ be the indices of one subpopulation of interest, and $B = b_1, b_2, \ldots, b_n$ be the other subpopulation of interest. We then exchange the posterior distributions in Equation 19 with the aggregated posterior:

$$q_\eta(z_a \mid C_A) q_\eta(z_b \mid C_B) = \left[\frac{1}{|A|}\sum_{a \in A} q_\eta\left(z_a \mid C_a\right)\right]\left[\frac{1}{|B|}\sum_{b \in B} q_\eta\left(z_b \mid C_b\right)\right]. \quad (21)$$

In this sampling procedure, a cell representation z_a (resp. z_b) is sampled given one randomly chosen cell in subpopulation A (resp. subpopulation B). Then, it is determined if $\left|\lambda_{a,b}^{g}\right| \geq \delta$ via an indicator function. The DE probability is estimated based on many samples.

Furthermore, by integrating over the batch variable s_n, we effectively compare cells as if they were in the same batch [15]. For genes, this is equivalent to computing

$$p\left(\left|\lambda_{a,b}^{g}\right| \geq \delta \mid C_a, C_b\right) \approx \sum_{s'} \int \mathbb{1}\{|[\Lambda(z_a, z_b, s', s')]^g| \geq \delta\} p(s') q(z_a \mid C_a) q(z_b \mid C_b) dz_a dz_b. \quad (22)$$

Here $p(s')$ is a uniform prior over batches. Every time we sample from the posterior, we decode the samples using the same batch indicator, averaging the DE probability over every possible batch indicator.

For proteins, we use the same framework, but define

$$\gamma_{a,b}^{t} \quad = \log_2(\mathbb{E}[\tilde{r}_{at} \mid \beta_{at}, v_{at}, z_a] + \epsilon) - \log_2(\mathbb{E}[\tilde{r}_{bt} \mid \beta_{bt}, v_{bt}, z_b] + \epsilon), \quad (23)$$

where the conditional expectation is equal to

$$\mathbb{E}[\tilde{r}_{at} | \beta_{at}, v_{at}, z_a] = \beta_{at}\alpha_{at}(1 - v_{at}). \qquad (24)$$

This is interpreted as the foreground mean if the cell was generated from the foreground, and zero otherwise. The added constant ϵ is a "prior count" that helps define the log fold change when $\mathbb{E}[\tilde{r}_{nt} \mid \beta_{nt}, v_{nt}, z_n] = 0$. For all analysis, we set $\epsilon = 0.5$. As with genes, we are interested in calculating $p(|\gamma_{a,b}^{t}| \geq \delta \mid C_a, C_b)$, where in this case we integrate with respect to the distribution

$$\prod_{i \in a,b} p\left(v_{it} \mid z_i\right) q(\beta_{it} \mid z_i, s_i) q(z_i \mid C_i). \qquad (25)$$

We consider features with a $\log(\mathrm{BF}) > 0.7$ as differentially expressed. This is roughly equivalent to calling features significant if the odds ratio (here equivalent to a Bayes factor) is greater than 2.

Finally, we use the posterior samples of $\lambda_{a,b}$ (resp. $\gamma_{a,b}$ for proteins) as the estimate of effect size for each gene (resp. protein). Specifically, we use the median of the samples, which is robust to outliers and is also the Bayes estimator under L_1 loss.

Denoised correlation matrix construction

We seek a feature-feature correlation matrix (e.g., gene-gene correlations, gene-protein cross-correlations) that summarizes biological variation, instead of technical variation. As totalVI explicitly models nuisance factors (for genes as well as proteins), we can query the model while controlling for this nuisance variation. Furthermore, because naive computations of correlations on denoised values (parameters of conditional distributions) were shown to induce spurious gene-gene correlations [33], we develop a novel sampling scheme that helps remove technical variation while avoiding such artifacts.

In order to ensure our correlation matrix does not include variation from the modeled technical factors, we perturb the data generating process to fix the library size ($\ell_n = 10000$) as well as incorporate the denoised protein expression conditional distribution. In particular, we compute a correlation matrix using samples from the distribution

$$p(\log w_n, \log \tilde{r}_n \mid C_{1:N}, \ell_{1:N}). \quad (26)$$

This is also a posterior predictive density whose samples are generated with ancestral sampling. As $\tilde{r}_n$ is zero-inflated, we add the same "prior count" before taking the logarithm. For this distribution, we sample ancestrally using the aggregated posterior

$$q_\eta(z_n, \beta_n \mid C_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} q_\eta(z_n \mid C_n) q_\eta(\beta_n \mid z_n, s_n), \quad (27)$$

One could in principle replace the aggregated posterior with the prior in case of analyzing dataset-wide correlations. However, this approach is more flexible as it can be applied to calculate the correlation matrix for a subpopulation $A = \{a_1, a_2, \ldots, a_m\}$, where A is the set of indices for the subpopulation, by conditioning the distribution on x_A and y_A.

The distinction between this procedure and those that induced spurious correlations is that the latter effectively estimates a correlation matrix using the expected value of the posterior predictive distribution, rather than estimating the correlation matrix of the posterior predictive distribution.

Out-of-batch generalization

totalVI learns a transformation from z_n and s_n to the parameters of the conditional distributions for each feature (decoder). In an out-of-batch prediction, we predict the expression of a cell (e.g., the mean of conditional distribution) given any of the other B observed batches s such that $s \neq s_n$. Here we describe a general way to sample posterior quantities for a cell while also "transforming" it into a different batch that was also observed for other cells [69]. Special cases of this have already been described in the protein imputation and differential expression sections. Consider, for

instance, the RNA counts in cell n and gene g. We can calculate posterior predictive samples of x_{ng} while conditioning on any arbitrary observed batch b. Then,

$$p\left(x_{ng}^{*} \mid C_n, s = b\right) \approx \int p_{\nu}\left(x_{ng}^{*} \mid z_n, s = b\right) q_{\eta}\left(z_n \mid C_n\right) dz_n. \qquad (28)$$

Furthermore, we can integrate over the choice of batch by sampling from

$$\sum_{b} p\left(x_{ng}^{*} \mid C_n, s = b\right) p(s = b), \qquad (29)$$

where $p(s)$ is a uniform prior over batches. We take the expected value of this particular distribution as batch-corrected, denoised gene expression data. This "transforming" can also be applied to other likelihood parameters like π_n.

CITE-seq experiment on mouse spleen and lymph node

Supplementary Table 1 shows a summary of the experimental design that generated the mouse spleen and lymph node CITE-seq datasets. Below, we describe in further detail how these datasets were collected and processed.

Cell preparation

Mice were group housed with enrichment in standard cages on ventilated racks at an ambient temperature of 26C and 40% humidity. Mice were kept in a dark/light cycle of 12 hours on and 12 hours off. Two female C57BL/6 (B6) mice at 5 weeks of age were euthanized using CO2. From each mouse, six lymph nodes were harvested, pooled in RPMI + 10% FBS media on ice, mechanically dissociated with a syringe plunger, and passed through a 70 μm strainer to generate a single cell suspension. Likewise, the spleen was harvested, placed in RPMI + 10% FBS media on ice, mechanically dissociated with a syringe plunger, and passed through a 70 μm strainer to generate a single cell suspension. For the spleen, red blood cells were lysed in Red Blood Cell Lysis Buffer (BioLegend #420302) following the manufacturer's protocol. All animal care and procedures were carried out in accordance with guidelines approved by the Institutional Animal Care and Use Committee at BioLegend, Inc.

Antibody panel preparation

We prepared panels containing either 111 antibodies (TotalSeq-A mouse antibody panel 1, BioLegend #900003217) or 208 antibodies (TotalSeq-A mouse antibody panel 2, BioLegend #900003218), which are enumerated in Supplementary Data. We performed a buffer exchange on each panel using a 50kDa Amicon spin column (Millipore #UFC505096) following the manufacturer's protocol to transfer antibodies into RPMI + 10% FBS. Spleen and lymph node cell suspensions were stained with different hashtag antibodies [22].

CITE-seq protocol and library preparation

The CITE-seq experiment was performed following the TotalSeq protocol with two slight modifications. First, the 10 minute centrifugation at 14,000g to remove antibody aggregates was conducted prior to buffer exchange. Second, cells were stained, washed, and resuspended in RPMI + 10% FBS to maintain viability. After staining, washing, and counting, 12,000 spleen cells and

12,000 lymph node cells were mixed and loaded into a single 10x lane. We followed the 10x Genomics Chromium Single Cell 3' v3 protocol to prepare RNA, antibody-derived-tag (ADT) and hashtag-oligo (HTO) libraries [70].

Sequencing and data processing

RNA, ADT, and HTO libraries were sequenced with an Illumina NovaSeq S1. Reads were processed with Cell Ranger v3.1.0 with feature barcoding, where RNA reads were mapped to the mouse mm10-2.1.0 reference (10x Genomics, STAR aligner [71]) and antibody reads were mapped to known barcodes (Supplementary Table 5). Hashtags were demultiplexed separately for each 10x lane with *HTODemux* in Seurat v3 using the kmeans function [34]. No read depth normalization was applied when aggregating datasets.

Additional datasets

We also used publicly available CITE-seq datasets from 10x Genomics. These included "10k PBMCs from a Healthy Donor - Gene Expression and Cell Surface Protein" (PBMC10k, [23]), "5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry)" (PBMC5k, [37]), and "10k Cells from a MALT Tumor - Gene Expression and Cell Surface Protein" (MALT, [24]). PBMC10k had 14,010 mean reads per cell for antibodies (5,816 median UMI counts per cell), while PBMC5k had 7,451 mean reads per cell for antibodies (2,752 median UMI counts per cell).

CITE-seq data pre-processing

For each dataset, after initial cell and gene filtering, we retained at least the top 4,000 highly variable genes (HVGs) as defined by the Seurat v3 method, merging HVGs from different batches when appropriate [34]. Dataset specific filtering is described below.

Spleen and lymph node

An initial cell filter removed cells expressing fewer than 200 genes. Cells labeled as either doublets or negative for hashtag antibodies by *HTODemux* were also removed. A protein library size filter retained cells with between 400 and 10,000 total protein UMI counts. We also filtered on the number of proteins detected. For cells stained with the 111 antibody panel, we removed cells with fewer than 90 proteins detected, while the cutoff was set to 170 for cells stained with the 208 antibody panel. Cells with a high percentage of UMIs from mitochondrial genes (15% or more of the cell's total UMI count) were removed. An initial gene filter removed genes expressed in 3 or fewer cells in any given batch. In addition to the top 4,000 HVGs selected by the Seurat v3 method, we retained genes that encode the proteins targeted by the 111 antibody panel. This resulted in 4,005 total genes. After all filters, the distribution of cells per dataset was: (SLN111-D1, 9,264 cells), (SLN111-D2, 7,564 cells), (SLN208-D1, 8,715 cells), (SLN208-D2, 7,105 cells). This is a total of 32,648 cells. Unless otherwise stated, we filtered out isotype control antibodies (9 total in the 208 panel) and hashtag antibodies. The protein CD49f was also removed due to having very low total UMI counts.

PBMC10k, PBMC5k, & MALT

For each of these datasets, we first removed doublets using DoubletDetection [72]. Cells with high mitochondrial content (percentage of UMIs from mitochondrial genes), high number of genes

detected, high UMI counts, and with fewer than 200 genes expressed were removed. Next, cells with outlier protein library size (on either end) were removed. Genes with expression in three or fewer cells were removed. Finally, the top 4,000 HVGs were retained. Dataset specific parameters are in Supplementary Table 6. In the case where the PBMC datasets are integrated, the 4,000 HVGs are selected by merging HVGs computed on each dataset separately as in the Seurat v3 method.

Posterior predictive checks and held-out metrics

Posterior predictive checks are useful to check the fit of Bayesian models. They work by comparing the observed data to posterior predictive samples from the model [25]. Much of the benchmarking done here was inspired by previous work done to benchmark the scHPF model [16]. We compared totalVI to factor analysis, which is a linear-Gaussian alternative to totalVI, and is easily extendable to multiple modalities as features are treated conditionally independent of the latent representation. Furthermore, we compared to scHPF, which received the concatenated RNA and protein count matrices as input. As a control, we also compared performance on RNA only to scVI [15]. Posterior predictive samples for totalVI and scVI were obtained by calling the *generate* function in the scVI package, which samples from the variational posterior distributions, and subsequently from the likelihood distributions given the posterior samples. We ran scVI with 20 latent dimensions and negative binomial conditional distribution in order to be consistent with totalVI. Factor analysis (FA) models were fit using the sklearn package [73] on the combined RNA and protein measurements using one of two normalization procedures. The first procedure consisted of transforming each value by $\log(\text{count} + 1)$. The second procedure consisted of log library size normalizing the RNA features and protein features separately. For example, considering only the RNA measurements for a cell, we normalized each cell to sum to 1 by dividing by the library size of RNA, multiplied by 10,000, added 1 to each value, and took a log transformation:

$$\tilde{x}_{ng} = \log\left(L\frac{x_{ng}}{\sum_g x_{ng}} + 1\right), \qquad (30)$$

where $L = 10000$. This process was then applied to the protein measurements. We refer to this type of normalization as *log library size normalization*, and for short, *log rate*. These normalization procedures are necessary as FA assumes a Gaussian distribution, so training on the raw data would lead to poor model fit. Posterior predictive samples for FA models were computed using the fitted parameters and posterior distribution derived in [74]. We note that normalization procedures were inverted so that FA posterior predictive samples were on the same scale as the raw data.

For each dataset, each model was trained on a train set comprising of 85% of the cells. An additional 5% of cells were held-out as a validation set for totalVI early stopping. The remaining 10% of cells were also held-out as a test set. For each model's posterior predictive samples (25 for each model) based on the train set, we calculated the coefficient of variation (CV) for each feature, and calculated the mean absolute error between the average CV and the observed raw data CV. Furthermore, we computed the Mann-Whitney U statistic (implementation in *scipy.stats.mannwhitneyu*) for each feature between the posterior predictive sample and the raw data. We averaged the statistic across all posterior predictive samples for each feature. We also used posterior predictive samples to assess generalization to unseen data. In this setup, we generated posterior predictive samples (150 for each model) conditioned on the test set. We

considered the mean absolute error between the observed held-out data and the posterior predictive mean.

Moreover, we computed a held-out calibration error [26] for each model based on the test set. For each cell n and gene g, let I_{ng} be the indicator that the observed value is contained in the interval of all posterior predictive samples. The calibration error for genes is then calculated as

$$\text{Cal}^{\text{RNA}} = \left(1 - \frac{1}{NG}\sum_{n}\sum_{g} I_{ng}\right)^2. \qquad (31)$$

The calibration error for proteins is computed separately following the same procedure.

Finally, for totalVI and scVI only, and for only the RNA data, we computed the held-out predictive log likelihood. In this metric, z_n and ℓ_n were sampled from the variational posterior for each cell n and the average negative conditional log likelihood, $-\log p(x_n \mid z_n, \ell_n, s_n)$ was computed. This is also called the reconstruction loss in the VAE literature. This is also an approximation of $-\log p(x_n \mid x_n, y_n, s_n)$, the negative predictive log likelihood. We note that we cannot compare the log likelihood of totalVI and scVI, which use discrete conditional distributions to factor analysis models, which use continuous conditional distributions.

We further evaluated model misfit through posterior dispersion indices [75]. This metric highlights cells that are not well explained by the model. This analysis is described in Supplementary Note 7.

Background decoupling benchmarking

We reported the totalVI background probability as the posterior predictive mean of π_{nt}, thus

$$p(\text{cell } n, \text{ protein } t \text{ is background}) = \mathbb{E}_{p(\pi_{nt} \mid x_n, y_n, s_n)}[\pi_{nt}], \qquad (32)$$

where the expectation is approximated using Monte Carlo integration. The totalVI foreground probability is one minus the background probability.

Observing protein background in empty droplets and non-expressing cell types

To observe different sources of protein background, we considered both empty droplets and cell types with known expression of surface markers. We defined empty droplets as non-cell barcodes from the SLN111-D1 dataset with between 20 and 100 RNA UMI counts (approximately 75,358 barcodes). We chose these criteria so that empty droplets were likely to represent ambient molecules rather than sequencing errors (with very low UMI counts) or cell debris (with higher UMI counts) [76]. To observe non-specific binding of antibodies, we considered B cells (which are known to express CD19 and CD20, but not CD28) and T cells (which are known to express CD28, but not CD19 or CD20). Using cell type annotations as described below, we grouped all high-quality, non-doublet B cell clusters (excluding plasma B cells), and alpha/beta T cell clusters (including all CD4, Treg, and CD8 T cell clusters). We observed that for these three proteins, both empty droplets and the non-expressing cell type contained protein background (non-zero protein counts) with varying degrees of overlap with the foreground signal of the expressing cell type. In

this text, we describe the protein counts of the non-expressing cell type above the counts in empty droplets as non-specific antibody binding, although we acknowledge there could be multiple sources of this cell-specific background (Supplementary Note 3).

Classification of cell type by marker proteins

We sought to evaluate totalVI against a Gaussian mixture model (GMM) at predicting major cell types by the foreground probability of commonly used surface markers. For these markers, protein counts were expected to come from the foreground component in some cell types and from the background component in others. For example, a high foreground probability for CD4 could be used as a positive predictor of CD4 T cells. We applied scikit-learn's *GaussianMixture* with default parameters to fit a GMM with two components to the log(protein counts + 1) for each protein for all cells in the SLN111-D1 dataset. We interpreted the posterior probability of the component with the higher mean as the foreground probability and that of the lower mean as the background probability. Restricting all cells to just those that fell into the categories of B cells or T cells as described above, we tested how well totalVI or a GMM could classify cell types based on commonly used protein surface markers. For each marker protein, we computed a receiver operating characteristic curve (ROC) (*sklearn.metrics*) by thresholding the totalVI or GMM foreground probability estimates, using manual cell type annotations as true labels (stratification and annotation described below). We reported the area under the ROC (ROC AUC). The cell type considered as the positive population was either B cells, T cells, CD4 T cells, or CD8 T cells depending on the marker. In tests considering each of these positive populations, all remaining cells among the B and T cells were considered the negative population. Marker proteins tested included, for B cells: CD19, CD45R-B220, CD20, I-A-I-E (MHC II); for T cells: CD5, TCRb, CD28, CD90.2; for CD4 T cells: CD4; for CD8T cells: CD8a, CD8b [77-80]. Although we are aware of documented exceptions to these markers appearing strictly on a single cell type (e.g., CD5 is expressed on a portion of B1 B cells), these exceptions are rare. In these cases where marker expression is not mutually exclusive, cell types can still be distinguished by the gradation in levels of the marker between cell populations. Thus, these exceptions do not negate the utility of these markers in broad cell type classification (which is apparent in both totalVI and GMM performance at this classification task).

GMM-based cutoff for protein foreground/background

As a baseline determination of a cutoff to distinguish cells with foreground or background protein expression, we used a GMM fit on all cells of the SLN111-D1 dataset for each protein as described above. The GMM-based cutoff between foreground and background was determined to be the protein expression level at which the GMM foreground probability (described above) was closest to 0.5.

Protein normalization using isotype controls

Although totalVI does not make use of isotype controls in its model of protein background, some CITE-seq studies include isotype control antibodies as negative controls to adjust for protein background. To compare totalVI to a method that uses isotype controls to normalize protein data, we applied two different normalization strategies prior to fitting a GMM and performing the classification task described above. First, we applied the normalization strategy used by Cumulus [32]:

$$\text{norm1}: y_{nt} \rightarrow \max\left(\log\frac{y_{nt}+1}{k_n^{(t)}+1}, 0\right), \qquad (33)$$

where y_{nt} is the observed UMI counts for protein t in cell n, and $k_n^{(t)}$ is the observed UMI counts of the corresponding isotype control for protein t in cell n. In the case where the corresponding isotype control for a given antibody is not present in the data, normalized expression is calculated as

$$\text{norm1}: y_{nt} \rightarrow \log(y_{nt}+1). \qquad (34)$$

Because this normalization method restricts normalized values to be non-negative, the resulting distribution might not be fit well by a GMM. We therefore applied a second normalization strategy as a modification to the Cumulus method that adjusts for the relative isotype control level but does not restrict the distribution of normalized values to be non-negative:

$$\text{norm2}: y_{nt} \rightarrow \log\frac{y_{nt}+1}{k_n^{(t)}+1}. \qquad (35)$$

If an isotype control is not present, *norm2* values are calculated as in Equation 34.
For the SLN208-D1 dataset, which contained a limited number of isotype control antibodies, we fit a GMM as described above to the log(protein counts + 1) *(GMM)*, to the Cumulus normalized values *(GMM norm1)*, and to the values normalized with the modified Cumulus method *(GMM norm2)*. We performed the same classification of cell types by marker proteins as described for the SLN111-D1 dataset, noting that the isotype control for CD28 (Syrian Hamster IgG) was not contained in the dataset.

Visualization and raw data normalization
For the SLN111-D1 dataset, we visualized the totalVI latent space in two dimensions using Scanpy's [56] implementation of the UMAP algorithm [81]. We applied log library-size normalization to the raw RNA counts as in Equation 30. All cells of the SLN111-D1 dataset are plotted (i.e., doublets were not removed).

Distribution of foreground probabilities
We observed the totalVI foreground probability for all proteins across all cells in the SLN111-D1 dataset (Supplementary Fig. 1e). The totalVI foreground probability tended to fall near zero or one. Measurements for which totalVI estimates a foreground probability near 0.5 are instances where the model is uncertain about whether the measurement is likely to be derived from foreground or background.

Distinguishing foreground and background in trimodal protein distributions
Despite using a two-component mixture, totalVI can decouple the background and foreground of proteins that have more than two modes globally. totalVI is capable of distinguishing foreground and background in this setting because the mixture is conditionally dependent on z_n, which allows the foreground and background expression modes to be defined locally in the latent space. For example, as has been reported using flow cytometry [82], CITE-seq data of peripheral blood mononuclear cells contains three modes of CD4 expression corresponding to CD4 T cells,

monocytes, and background. totalVI detected that both CD4 T cells and monocytes had foreground expression of CD4, while the CD4 expression of the remaining cells was attributable to background expression.

Denoised protein expression
Denoised protein expression was calculated as previously described. B cells and T cells were defined by annotations, as described above.

RNA-protein correlation analysis
Evaluation of correlation calculation with permuted features
Using totalVI, we aimed to calculate a correlation matrix between all RNA and protein features free from nuisance variation such as sequencing depth and protein background. We took care to avoid the naive calculation of correlations directly between denoised features, noting that a recent study reported false positive correlations in smoothed scRNA-seq data [33]. Instead, we developed a novel sampling method for the calculation of denoised feature correlations that removes nuisance variation while avoiding imputation-induced artifacts (described above).

To evaluate whether totalVI could calculate a denoised feature correlation matrix without introducing spurious relationships in the data, we permuted the expression of a set of genes to serve as a negative control. To create this set of negative control genes from the SLN111-D1 dataset, we selected the 100 genes with highest mean expression that were not already among the top highly variable genes used in the model. We randomly permuted the counts of these genes within each cell, rendering these genes independent of all other gene and protein features. After concatenating the SLN111-D1 dataset with the permuted gene expression for all cells, we ran the totalVI model.

We then calculated Pearson and Spearman correlations between features using three methods, referred to here as raw, naive totalVI, and totalVI correlations. Raw correlations were calculated between log library-size normalized RNA (Equation 30) and log(protein counts + 1). Naive totalVI correlations were calculated between totalVI denoised RNA and totalVI denoised proteins. totalVI correlations were calculated by sampling denoised RNA and denoised protein values from the posterior (as described above).

We observed the correlations between all RNA and protein features as well as the 100 additional genes whose expression was randomly permuted. By comparing the raw correlations with denoised correlations, we observed whether the method of denoising could maintain the relationship between these permuted genes and other features, which, in expectation, are independent from each other. Here, we highlighted the correlations between all proteins and the randomly permuted genes, whose correlations are expected to be near zero.

Correlations of RNA-protein pairs
We calculated a feature correlation matrix for the SLN111-D1 dataset using either the totalVI sampling method or by calculating raw correlations as described above. The resulting feature correlation matrices for both Pearson and Spearman correlations were subset to each protein and

its encoding RNA for all proteins with a unique encoding RNA in the dataset (i.e., excluding RNA with multiple isoforms such as Ptprc). It is worth noting that the totalVI model has no explicit information about the relationship between RNA-protein pairs, such that any correlation learned by the model is not predetermined by known RNA-protein relationships.

Integration of multiple datasets

We compared totalVI's integration performance to that of Scanorama[35], Seurat v3 [34], and Harmony [36]. The two former methods, like totalVI, produce both integrated expression values and integrated low-dimensional cell representations. The input to both Scanorama (*scanorama.correct*), Seurat v3 (*FindIntegrationAnchors*, *IntegrateData*) methods was a normalized matrix of concatenated genes and proteins. Genes were subset to the same subset used as input to totalVI. The RNA counts of this matrix were normalized following standard log library size normalization (Equation 30). For proteins, we used a $y \to \log(y+1)$ transformation. Finally, we standard scaled each feature. Harmony (*harmonypy*) received latent spaces for each dataset computed with PCA on the concatenated, normalized, and scaled datasets. All methods were run with default parameters. We compared the performance of the methods using the following metrics:

Latent mixing metric

The latent mixing metric measures how well the latent cell representations are mixed between batches relative to the global frequency of batches. First, a cell-cell similarity matrix is computed from a latent representation of cells. Next, select 100 cells uniformly at random, and calculate the frequency of batches represented in each cell's 100 nearest neighbors. Let $p_i^{(n)}$ be the frequency of batch i in the 100 nearest neighbors of cell n. Let q_i be the global frequency of batch i. Compute the negative relative entropy between the frequency of observed batches in the neighborhood, and the global frequency of batches:

$$\mathrm{KL}\left(p^{(n)} \parallel q\right) = \sum_{i=1}^{B} p_i^{(n)} \log \frac{p_i^{(n)}}{q_i} \qquad (36)$$

Repeat this 50 times and return the average negative relative entropy. This is conceptually similar to the entropy of mixing that has been used in other studies [83].

Measurement mixing metric

The measurement mixing metric describes how well the high-dimensional measurements are batch corrected, and for each feature, is related to the Mann-Whitney U statistic. Consider one feature in the batch-corrected data matrix placed in rank order. Let R_1 be the sum of the ranks of the cells in batch 1 and N_1 be the number of cells in batch 1. Define $U_1 = R_1 - \frac{N_1(N_1+1)}{2}$. Similarly, compute U_2 for batch 2 and return $\min(U_1, U_2)$. Higher values of this metric indicate better mixing within that feature. This metric could not be applied for Harmony, which only produces an integrated latent representation.

Feature retention metric

The feature retention metric describes how spatial autocorrelation of both RNA and protein change when comparing cells from an integrated latent representation to a latent representation derived

from each batch separately. Lower values of this metric indicate that the integration procedure reduced the localization of feature expression, indicating some degree of random mixing. We calculate it as follows. For two batches and a particular integration method, we calculate Z_1 and Z_2, the latent representations of the cells of batch 1 and batch 2, respectively. The latent space computation of the individual batches was chosen to closely match the integration method (see below). We also calculate an integrated latent representation of both batches $Z^\top = [\tilde{Z}_1 \;\; \tilde{Z}_2]$. Let $D_1 = [X_1 \;\; Y_1]$ be the combined RNA and protein batch 1 in which RNA is library size log normalized and proteins are log-transformed. Let $\mathbb{E}[H(D_1, Z_1)]$ be the expected feature autocorrelation score as calculated by Hotspot [84]. Furthermore, let $\mathbb{E}\left[H\left(D_1, \tilde{Z}_1\right)\right]$ be the analogous quantity calculated using the latent cell representations of batch 1 subsetted from the joint, integrated representation. The feature retention metric is calculated as $\frac{1}{2}\sum_i^2 \mathbb{E}\left[H\left(D_i, \tilde{Z}_i\right)\right] - \mathbb{E}[H(D_i, Z_i)]$. In the case of totalVI union, features were intersected to compute this metric.

For Scanorama, we define Z_1 and Z_2 to be a 100-dimensional matrix produced with principal components analysis (PCA), which is the same dimension reduction used in the integration method. For Seurat v3, we similarly use PCA to reduce D_1 and D_2 to 30 dimensions, the same number of dimensions used for integration. The input to PCA was the same as the input for the respective method, except for Scanorama, where we additionally L_2 normalized each cell, because this step is done automatically by Scanorama's *correct* method.

Clustering metric

The clustering metric quantifies the extent to which clusters defined on the unintegrated latent spaces are preserved in the integrated latent space. Using the same notation as before, we compute for each method, clusters based on Z_1 and Z_2, individually. Clusters were inferred using the standard Scanpy workflow: computing a neighbors graph, and running the Leiden [85] algorithm, with default parameters. Next, the silhouette coefficient $\mathcal{S}$ was computed for every cell with respect to its latent representation and cluster label: $\mathcal{S}(Z_1), \mathcal{S}(Z_2), \mathcal{S}\left(\tilde{Z}_1\right), \mathcal{S}\left(\tilde{Z}_2\right)$. Finally, a score for each dataset was defined as $\mathbb{E}\left[\mathcal{S}\left(\tilde{Z}_i\right) - \mathcal{S}(Z_i)\right]$. The final score was averaged across each dataset. Thus, lower scores suggest clusters are not preserved as well in the integrated latent space. We emphasize that this metric can only be taken as a proxy for cell type preservation, which requires "ground truth" cell type labels, or well-established datasets -- none of which exist for CITE-seq.

Missing protein imputation

For Seurat v3, we imputed proteins based on mutual nearest neighbors in the RNA data using the *FindTransferAnchors* and *TransferData* functions. Again, RNA data were log library size normalized. Proteins were not normalized as input to Seurat. For totalVI, after fitting the model, cells from the batch with held-out proteins were decoded conditioned on being in the batch with observed protein data. Note, we did not correct for background in this analysis since the comparison is to the observed data. We used the root mean squared error of values on the log scale to assess imputation accuracy. To produce error bars, we ran totalVI 30 times, resampling the dataset into the train/validation sets (validation used for early stopping), computing the mean and 95% confidence interval. For the PBMC datasets, we compared to cTP-net [38], which is a neural network that was pre-trained on specific CITE-seq datasets from human cells, with no option to train a new dataset. The inputs to cTP-net were the log-normalized RNA data. cTP-net did not

provide predictions for CD127, CD15, CD25, PD-1, or TIGIT. To the best of our knowledge, neither of the PBMC datasets used in this study were used to train the pre-trained cTP-net model. Thus, a direct comparison of the results to those of totalVI or Seurat v3 is not straightforward.

Stratification of cells in SLN-all

We stratified cells of the mouse spleen and lymph node based on the SLN-all dataset (totalVI-intersect model fit as described above). We clustered cells in the totalVI latent space with Scanpy's implementation of the Leiden algorithm at resolution 1, resulting in 32 clusters [56, 85]. We repeated this approach to sub-cluster cells, finding a total of 43 clusters. We used Vision [49] with default parameters for data exploration, including its implementation of the Wilcoxon rank sum test, to identify cluster markers. Clusters were manually annotated based on a curated list of cell type markers (Supplementary Table 4). Clusters annotated as doublets, low quality cells (e.g., high percentage of UMI counts from mitochondrial genes), or cells undergoing the cell cycle were removed from further analysis. Again, we visualized the totalVI latent space in two dimensions using Scanpy's implementation of the UMAP algorithm. These annotations were also consistent with the latent space derived with totalVI-union (Supplementary Fig. 7).

Differential expression analysis

The Welch's t-test and Wilcoxon rank-sum test for each differential expression scenario were run on protein features (log-transformed) using the Scanpy library, which produces adjusted p-values corrected for multiple testing by the Benjamini-Hochberg procedure [86]. Both tests are two-sided. A protein was considered to be differentially expressed if the adjusted p-value was less than 0.05. Each application of totalVI differential expression tests to a dataset requires a trained totalVI model. For each dataset used in DE analysis, all cells were included to train the model. Throughout, we used our manual annotations from the SLN-all totalVI-intersect model run. The cells in nuisance clusters (described in previous section) were removed before running totalVI differential expression functions.

In a given totalVI differential expression test, we identified cell type markers by first filtering features for significance (log Bayes factor > 0.7), and then sorting by the median log fold change. We only retained genes with non-zero UMI counts in at least 10% of the subset of cells.

In the comparison to scVI gene Bayes factors, each method was trained independently on the SLN111-D1 dataset. We ran scVI with 20 latent dimensions and negative binomial conditional distribution to be consistent with totalVI. Differential expression of genes in scVI was computed using the same LFC-based method, which is implemented in the *scvi-tools* package. In reproducibility benchmarking, totalVI was trained independently on the replicates.

In the test between ICOS-high Tregs and CD4 conventional T cells, we used the same totalVI-intersect model fit that was used to manually annotate the cells. In this test, we expected CD73, CD357 (GITR), CD122, and CD5 to be upregulated (positives) in ICOS-high Tregs relative to conventional CD4 T cells [87-90]. The list of putative negatives included I-A/I-E (MHC II), IgD, CD19, CD8b, and CD8a, which have no expected expression in either of these cell types.

DE on imputed proteins

In one totalVI model fit, SLN111-D1 and SLN111-D2 were integrated with the proteins of SLN111-D2 held out. In the second totalVI model fit, these two datasets were integrated with all data. In testing differential expression of proteins, and for each model fit, we conditioned on SLN111-D1. This is an application of Equation 22, except that the prior $p(s')$ is 1 if $s' =$ SLN111-D1 and 0 otherwise.

Archetypal analysis

This analysis was performed on the SLN-all totalVI-intersect model run. As z_n is distributed as logistic normal, the latent space is then constrained to the probability simplex (i.e., each z_n is non-negative and sums to one). Archetypes correspond to vertices of the totalVI latent space, which means they can be represented by the identity matrix I_d, where d is the number of latent dimensions (20 in all experiments). In this setting, the latent space is the 19-dimensional standard simplex.

We first identified and removed four archetypes from further interpretation that suffered from inactivity (a known issue in training VAEs) [91]. For the remaining 16 latent dimensions, we decoded the archetypes to obtain denoised RNA and protein archetypal expression profiles, all conditioned on batch 0 (the SLN111-D1 experiment). We then computed denoised RNA and protein expression profiles for all cells in SLN-all, conditioned on SLN111-D1. To derive signatures for each archetype, we computed the mean and standard deviation of each feature in the denoised RNA and protein expression matrices (without the archetypes) and standard scaled the archetypal profiles with respect to this mean and standard deviation. We refer to this quantity as the archetype score. The top features for each archetype were those with an archetype score greater than 2. The distance to the archetype is computed as the Manhattan distance from each cell's latent representation to the archetype. The distances per archeytpe were scaled into the range $[0,1]$.

B cell analysis

For this analysis, we used the totalVI-intersect model fit on the SLN-all dataset as described above. The SLN-all dataset was filtered to include all high-quality, non-doublet clusters annotated as B cells (excluding plasma B cells), resulting in 15,560 cells.

Calculation of signature scores

Gene signature analysis was conducted using Vision [49] with default parameters. Gene signatures, including interferon response signatures, were downloaded from MSigDB gene sets [92]. Signature scores were calculated on all cells in the SLN-all dataset based on cell similarities defined by the totalVI latent space.

Identification of transitional and mature B cell feature modules

totalVI Spearman correlations between all features were calculated separately within the transitional B cell cluster and the mature B cell cluster. Features were subset by the following method. From a one-vs-one DE test between transitional and mature B cells, we selected the top ten marker genes and top three marker proteins for each cluster (as described above). We added to this list the four features most highly correlated with each differentially expressed feature within

its respective cluster. This resulted in a list of both transitional and mature features which we used to subset the full feature correlation matrix. Features were hierarchically clustered separately for transitional and mature B cells using Seaborn's *clustermap* with default parameters.

When plotting totalVI expression of each feature as a function of $1 - Z_{16}$, each feature was standard scaled and smoothed with a loess curve. Spearman correlations were calculated between each feature and $1 - Z_{16}$. The p-values of these correlations were all significant (BH-adjusted p-value < 0.001).

Data availability

The data discussed in this manuscript (SLN-all) have been deposited in NCBI's Gene Expression Omnibus [93] and are accessible through GEO Series accession number GSE150599 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150599). Processed data are also available in the reproducibility GitHub repository (https://github.com/YosefLab/totalVI_reproducibility). The SLN-all dataset processed with totalVI can be explored interactively with Vision at http://s133.cs.berkeley.edu:9000/Results.html. Public datasets were downloaded from 10x Genomics (PBMC5k: https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_protein_v3; PBMC10k: https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3; MALT: https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/malt_10k_protein_v3). Mouse mm10 reference was downloaded from 10x Genomics.

Code availability

The code to reproduce the results in this manuscript is available at https://github.com/YosefLab/totalVI_reproducibility and has been deposited at https://doi.org/10.5281/zenodo.4330368 [94]. The reference implementation of totalVI is available via the *scvi-tools* package at https://github.com/YosefLab/scvi-tools.

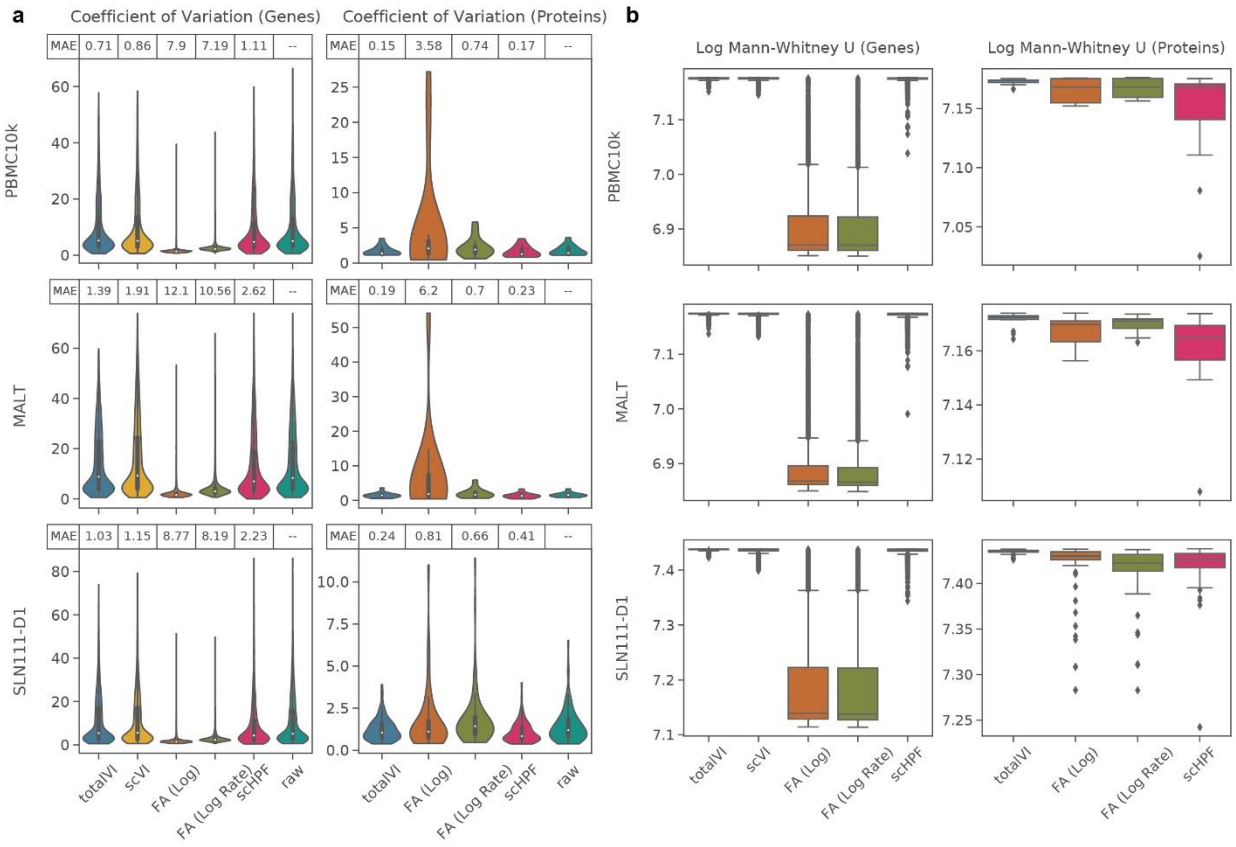
Additional references

59. Svensson, V., Gayoso, A., Yosef, N. & Pachter, L. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics* (2020) doi:10.1101/737601.
60. Wang, C. & Blei, D. M. A general method for robust Bayesian modeling. *Bayesian Analysis* (2018) doi:10.1214/17-BA1090.
61. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* (2020).
62. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877 (2017).
63. Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K. & Winther, O. Ladder variational autoencoders. in *Neural information processing systems* (2016).

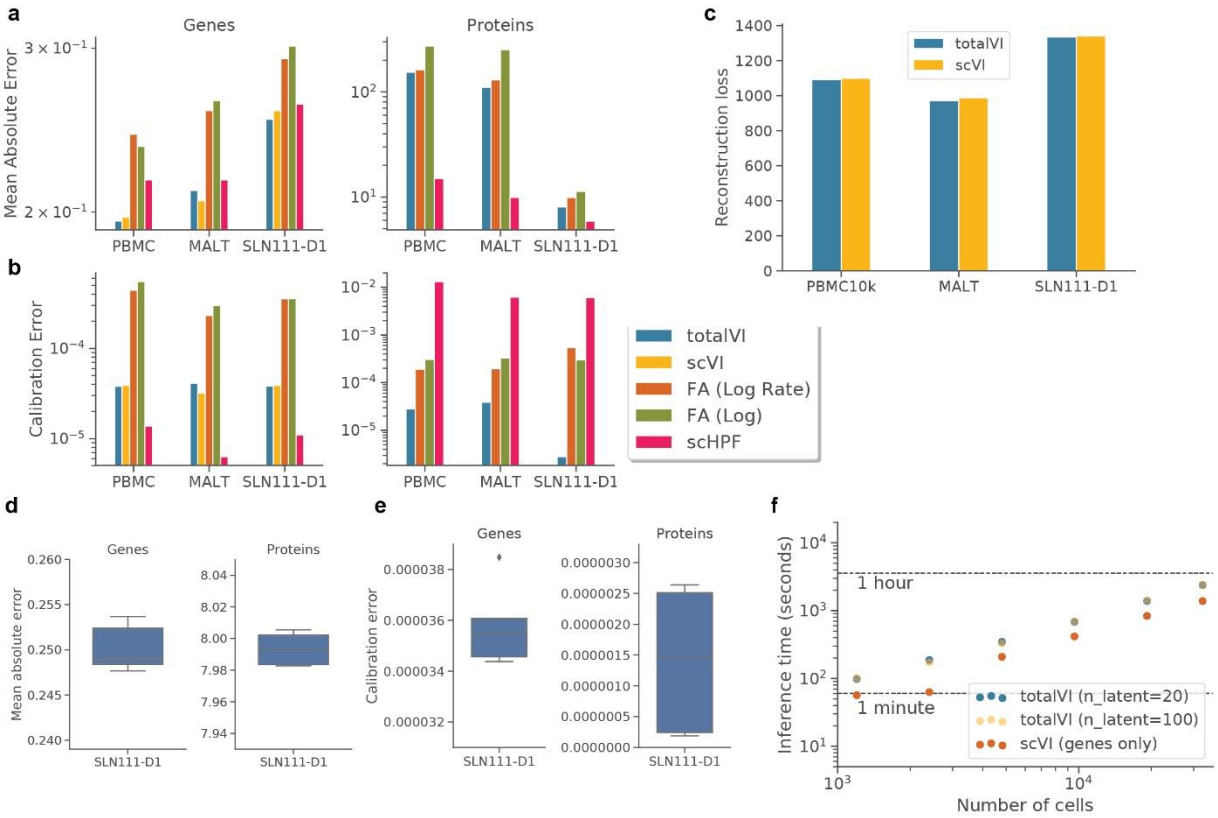
64. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. in *International conference on learning representations* (2015).
65. Lopez, R. *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. in *ICML workshop in computational biology* (2019).
66. Mattei, P. A. & Freiisen, J. Miwae: Deep generative modelling and imputation of incomplete data sets. in *International conference on machine learning* (2019).
67. Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. & Wortman, J. Learning bounds for domain adaptation. in *Advances in neural information processing systems* (2008).
68. Ganin, Y. *et al.* Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* **17**, 2096–2030 (2016).
69. Lotfollahi, M., Naghipourfar, M., Theis, F. J. & Wolf, F. A. Conditional out-of-sample generation for unpaired data using trVAE. *arXiv* (2019).
70. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* (2017) doi:10.1038/ncomms14049.
71. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) doi:10.1093/bioinformatics/bts635.
72. Gayoso, Adam, Shor, Jonathan, Carr, Ambrose J., Sharma, Roshan, Pe'er, Dana (2018, July 17). DoubletDetection (Version v2.4). Zenodo. doi: 10.5281/zenodo.2678041
73. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* (2011).
74. Bishop, C. M. *Pattern Recogniton and Machine Learning*. (2006).
75. Kucukelbir, A., Wang, Y. & Blei, D. M. Evaluating Bayesian models with posterior dispersion indices. in *International Conference on Machine Learning* (2017).
76. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology* **20**, 63 (2019).
77. Lai, L., Alaverdi, N., Maltais, L. & Morse, H. C. Immunophenotyping mouse cell surface antigens: Nomenclature and immunophenotyping. *The Journal of Immunology* (1998).
78. Watts, C. Capture and processing of exogenous antigens for presentation on MHC molecules. *Annual Review of Immunology* **15**, 821–850 (1997).
79. Uchida, J. *et al.* Mouse CD20 expression and function. *International Immunology* (2004) doi:10.1093/intimm/dxh009.
80. Hünig, T., Beyersdorf, N. & Kerkau, T. CD28 co-stimulation in T-cell homeostasis: a recent perspective. *ImmunoTargets and Therapy* **4**, 111 (2015).
81. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv* (2018).
82. Filion, L. G., Izaguirre, C. A., Garber, G. E., Huebsh, L. & Aye, M. T. Detection of surface and cytoplasmic CD4 on blood monocytes from normal and HIV-1 infected individuals. *Journal of Immunological Methods* **135**, 59–69 (1990).
83. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36**, 421–427 (2018).
84. DeTomaso, D. & Yosef, N. Identifying informative gene modules across modalities of single cell genomics. *bioRxiv* (2020) doi:10.1101/2020.02.06.937805.
85. Traag, V., Waltman, L. & Eck, N. J. van. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, (2019).

86. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
87. Zhao, H., Liao, X. & Kang, Y. Tregs: Where we are and what comes next? *Frontiers in Immunology* (2017) doi:10.3389/fimmu.2017.01578.
88. Roncarolo, M.-G. & Gregori, S. Is FOXP3 a bona fide marker for human regulatory T cells? *European Journal of Immunology* **38**, 925–927 (2008).
89. Fontenot, J. D., Rasmussen, J. P., Gavin, M. A. & Rudensky, A. Y. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nature Immunology* **6**, 1142–1151 (2005).
90. Sprouse, M. L. *et al.* High self-reactivity drives T-bet and potentiates Treg function in tissue-specific autoimmunity. *JCI Insight* **3**, 1–14 (2018).
91. Burda, Y., Grosse, R. & Salakhutdinov, R. Importance weighted Autoencoders. in *International conference on learning representations* (2016).
92. Liberzon, A. *et al.* Databases and ontologies Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
93. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).
94. Gayoso, A. and Steier, Z. (2020, December 18). YosefLab/totalVI_reproducibility: totalVI reproducibility v0.3 (Version v0.3). Zenodo. https://doi.org/10.5281/zenodo.4330368.

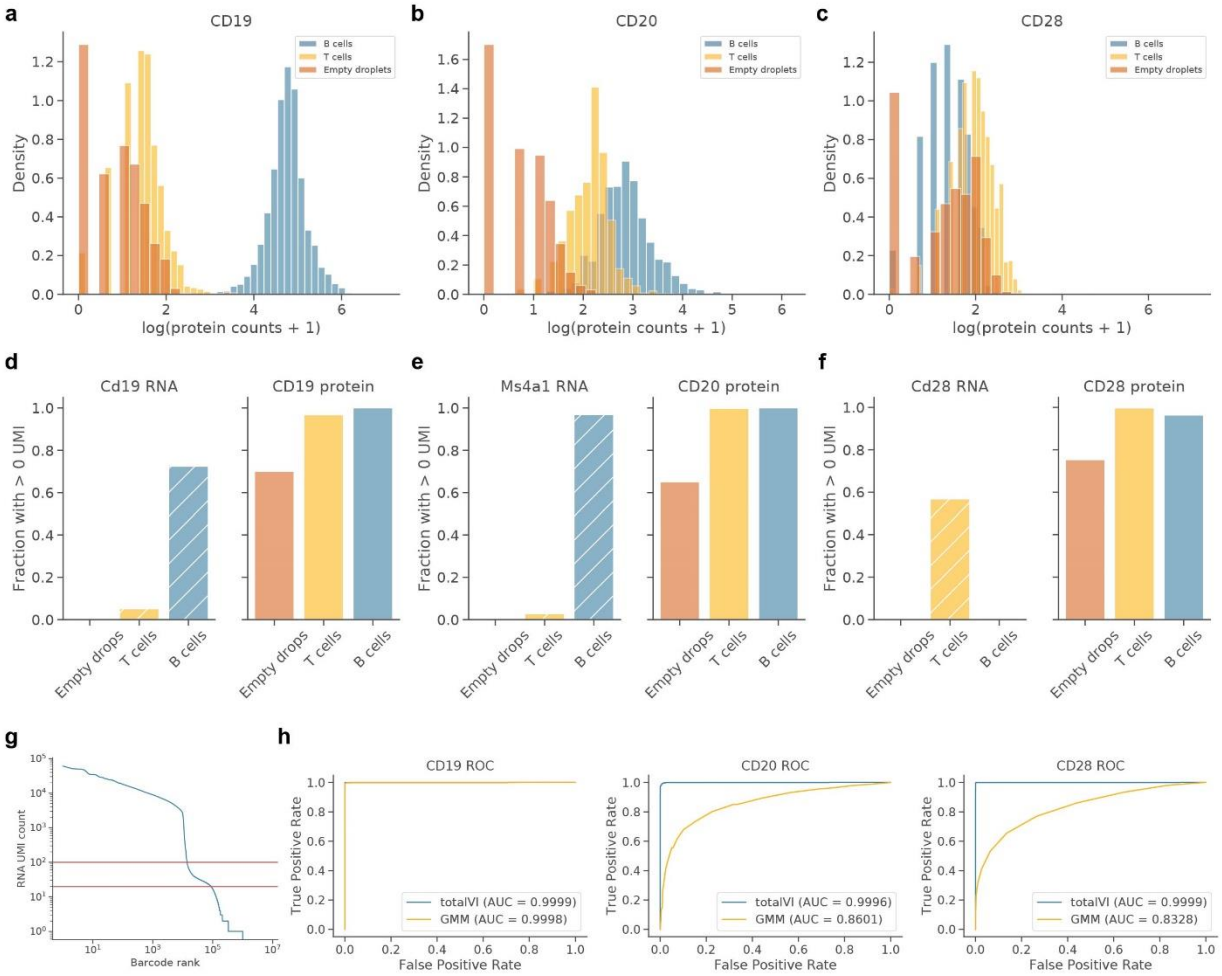
Extended data figures



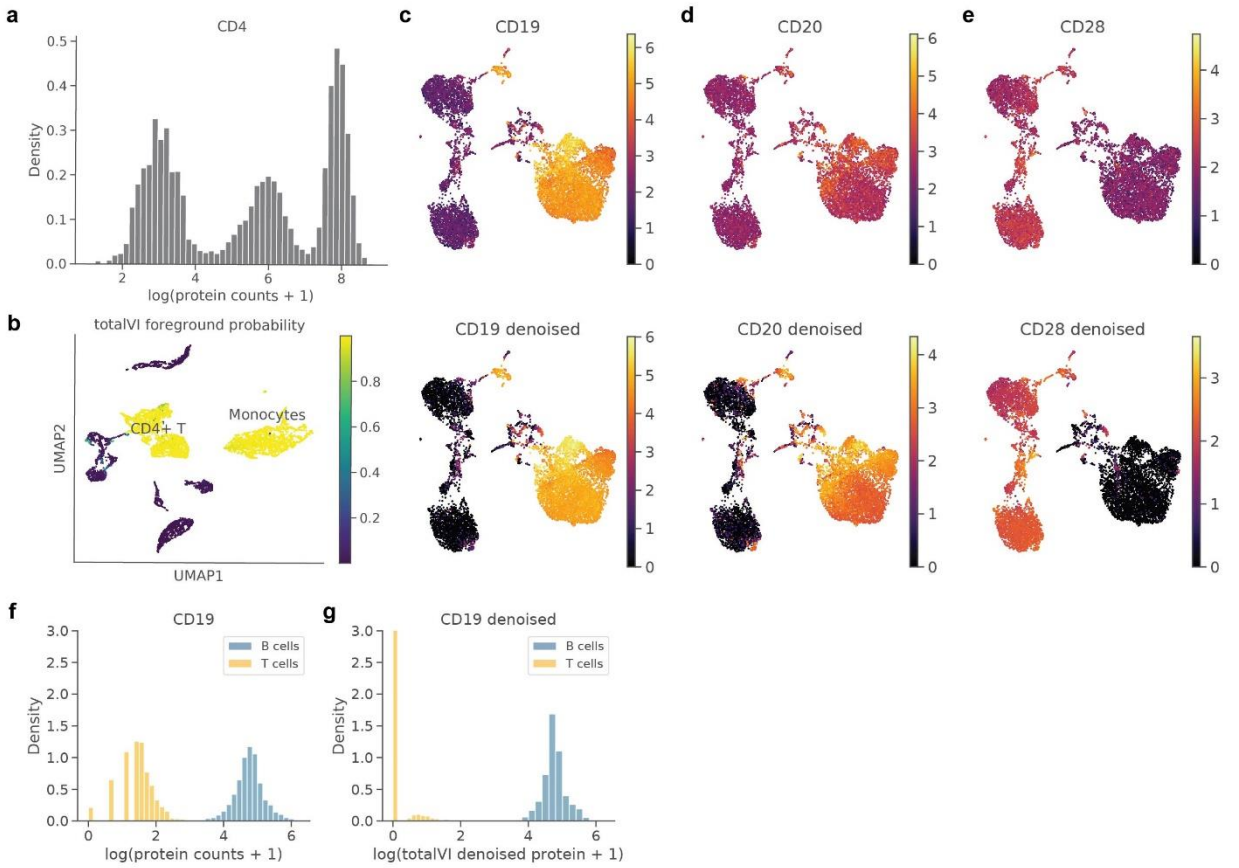
Extended Data Fig. 1 Evaluation of totalVI model. a, Posterior predictive check of coefficient of variation (CV) of genes and proteins. For each of the PBMC10k, MALT, and SLN111-D1 datasets and for each model (totalVI, scVI, factor analysis with normalized input, scHPF) the average coefficient of variation from posterior predictive samples was computed for each feature. Violin plots summarize the distribution of CVs for genes and proteins. Mean absolute error (MAE) between raw data CVs and average posterior predictive CV are reported. **b**, For each gene and protein, the Mann-Whitney U statistic between posterior predictive samples and observed data averaged over samples. Shown are boxplots of this statistic for each set of features (genes and proteins), model, and dataset (n=4000 genes across datasets and n=14 proteins for PBMC10k and MALT, n=110 proteins for SLN111-D1). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. Higher is better.



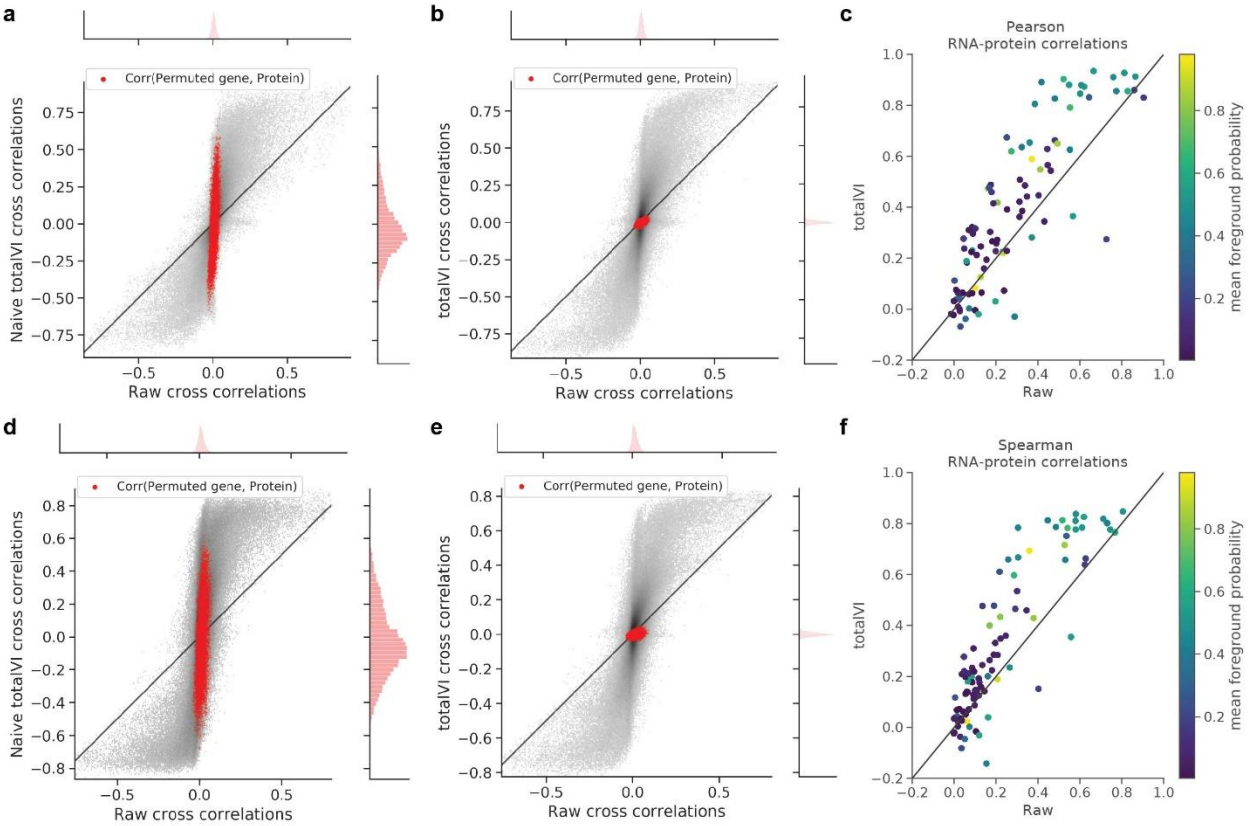
Extended Data Fig. 2 Evaluation of totalVI model (continued). a, Mean absolute error (MAE) between held out data and posterior predictive mean separated by genes and proteins for each model and dataset. **b**, Calibration error of held-out data reported separately for genes and proteins. **c**, Held-out reconstruction loss of RNA for scVI and totalVI. **d, e**, Stability of held-out results (*n*=5 initializations) for totalVI on SLN111-D1. Metrics displayed are the (**d**) Held out MAE, and (**e**) held out calibration error. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **f**, Inference time for totalVI and scVI across cells randomly subsampled to different levels from SLN-all. scVI was run with only genes. totalVI was applied with 20 latent dimensions and 100 latent dimensions.



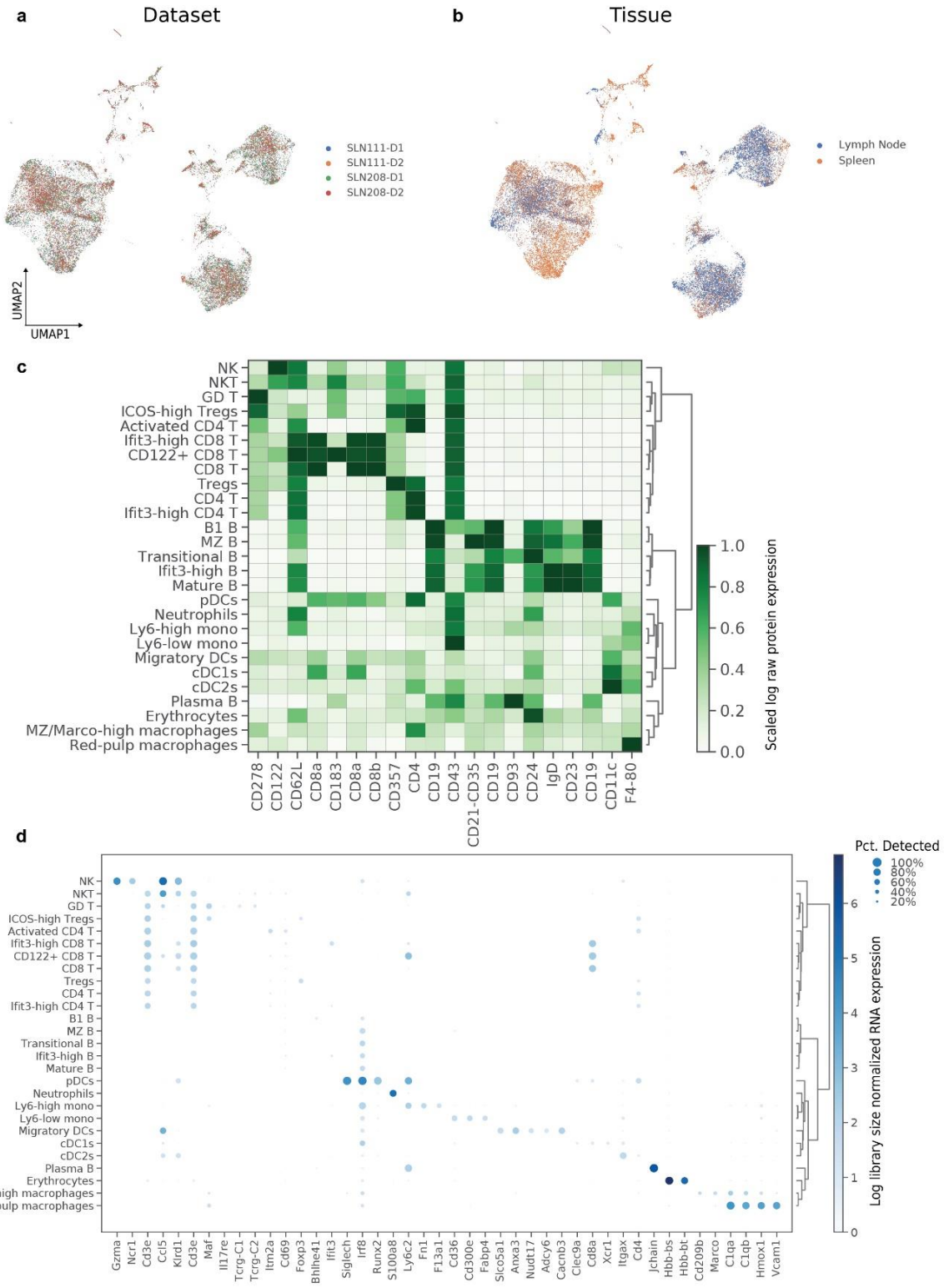
Extended Data Fig. 3 Protein background in cells and empty droplets. a-c, Histogram of log(protein counts + 1) in the SLN111-D1 dataset for B cells, T cells, and empty droplets (Methods) for CD19 (**a**), CD20 (**b**), and CD28 (**c**). d-f, Fraction of empty droplets, B cells, or T cells with > 0 UMIs detected for a given RNA (left, hatched) or protein (right, solid). RNA/proteins displayed are *Cd19*/CD19 (**d**), *Ms4a1*/CD20 (**e**), and *Cd28*/CD28 (**f**). **g**, Barcode rank plot for all barcodes detected in the SLN111-D1 dataset. Red lines at 20 and 100 RNA UMI counts indicate the lower and upper bounds, respectively, used to define empty droplets in (**a-f**). **h**, Performance of totalVI and a Gaussian mixture model (GMM) fit on all cells for each protein of the SLN111-D1 dataset to classify cell types by marker proteins (Methods). Receiver operating characteristic (ROC) curves shown for CD19 (B cells), CD20 (B cells), or CD28 (T cells). Area under the receiver operating characteristic curve (ROC AUC score) was calculated using as input either the totalVI foreground probability or GMM foreground probability where the indicated cell type was the positive population out of all B and T cells.



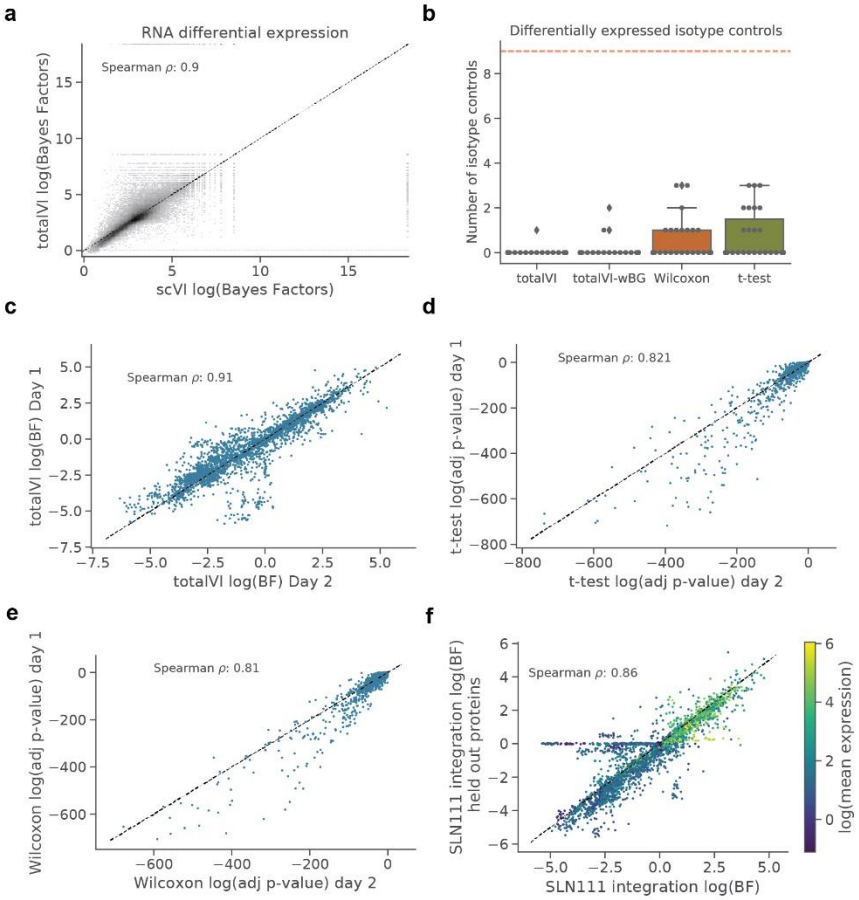
Extended Data Fig. 4 totalVI decouples foreground and background for trimodal protein distributions and denoises protein data. a, b, CD4 protein expression in the PBMC10k dataset. (**a**) Trimodal distribution of log(protein counts + 1). (**b**) UMAP plot of the totalVI latent space colored by totalVI foreground probability. **c-e**, UMAP plots of the totalVI latent space for the SLN111-D1 dataset. Plots are colored by log(protein counts+1) (top) and log(totalVI denoised protein+1) (bottom) for CD19 (**c**), CD20 (**d**), and CD28 (e). **f, g**, Distributions of log(protein counts + 1) (**f**) and log(totalVI denoised protein + 1) (**g**) for CD19 protein in B and T cells. y-axis is truncated at 3.



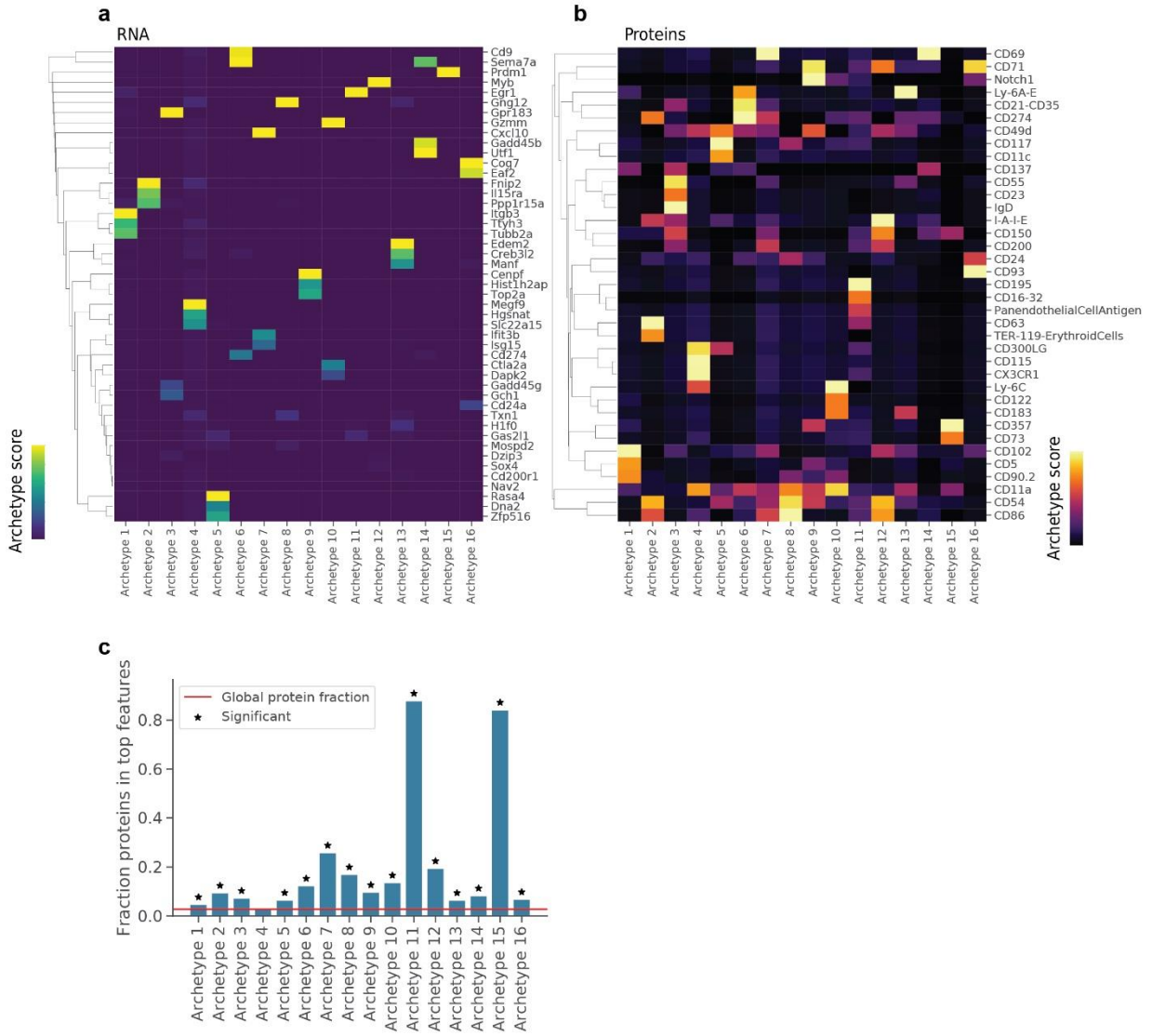
Extended Data Fig. 5 RNA-protein correlations. a, b, 2d density plots of Pearson correlations between all RNA and protein features in the SLN111-D1 dataset as well as 100 additional genes whose expression was randomly permuted. Correlations between all proteins and the randomly permuted genes are colored in red. Raw correlations were calculated between log library-size normalized RNA and log(protein counts + 1). (**a**), Naive totalVI correlations were calculated between totalVI denoised RNA and totalVI denoised proteins. (**b**), totalVI correlations were calculated between denoised RNA and proteins sampled from the posterior (Methods). **c**, Pearson correlations between each protein and its encoding RNA for all proteins with a unique encoding RNA, colored by the mean probability foreground of the protein across all cells. totalVI correlations were calculated as in (**b**) and raw correlation were calculated as in (**a, b**). **d-f**, Same as (**a-c**), but for Spearman correlations.



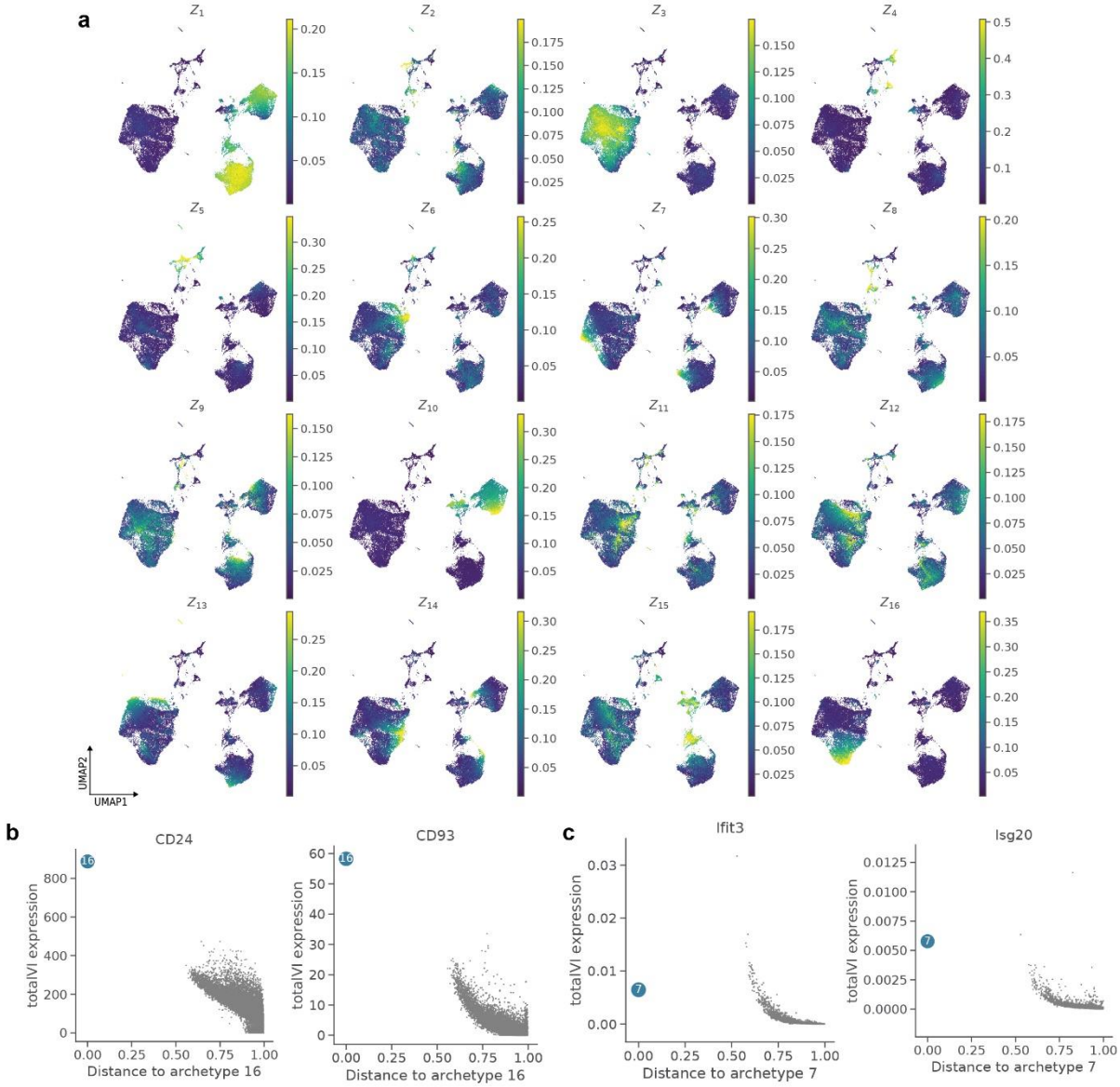
Extended Data Fig. 6 Integration of SLN-all with totalVI-intersect. a, b, UMAP plot of SLN-all colored by (**a**) dataset, and (**b**) tissue. **c**, Heatmap of proteins used for annotation. Proteins (columns) are log(protein counts + 1) scaled by column for visualization. **d**, Dotplot of RNA markers used for annotation. RNA is log library size normalized.



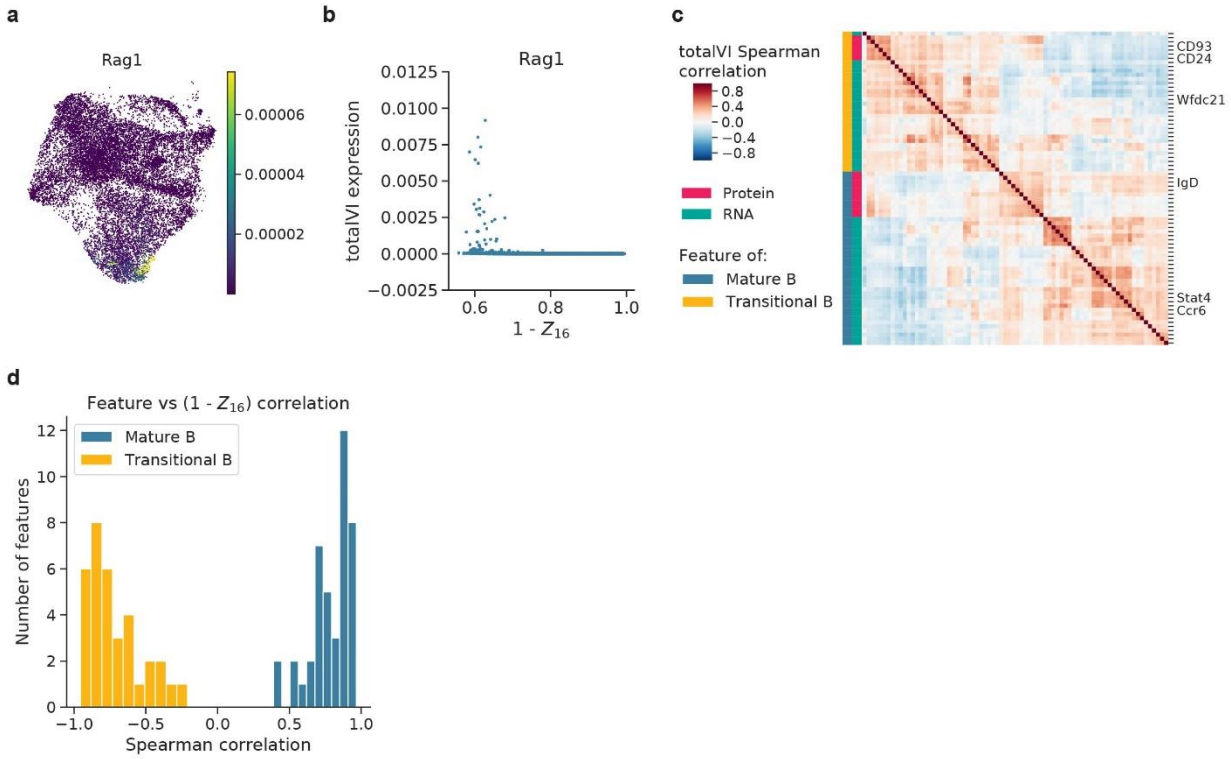
Extended Data Fig. 7 Differential expression analysis. a, 2d density plot of totalVI and scVI log Bayes factors for genes. Bayes factors were computed for each gene in one-vs-all tests on the SLN111-D1 dataset. **b**, Number of isotype controls called differentially expressed in one-vs-all tests (n=27) for totalVI, totalVI-wBG (totalVI test without background removal), Wilcoxon rank-sum, and t-test. Tests were applied to SLN208-D1, for which isotype controls were retained. Box plots indicate the median (center lines), interquartile range (hinges), whiskers at 1.5x interquartile range. Red dashed line indicates the maximum number of isotype controls. **c-e**, Significance level (Bayes factors for totalVI, adjusted p-value for frequentist tests) for proteins in one-vs-all tests computed on SLN111-D1 and SLN111-D2 for each of (**c**) totalVI, (**d**) t-test, (**e**) Wilcoxon. **f**, Bayes factors for proteins in one-vs-all tests computed on the SLN111 datasets integrated with and without the SLN111-D2 proteins held-out. Differential expression tests for both model fits were conditioned on SLN111-D1. Bayes factors are colored by the average protein expression from SLN111-D1.



Extended Data Fig. 8 Interpreting totalVI latent dimensions with archetypal analysis. a, b, Heatmap of top (**a**) gene and (**b**) protein features for each archetype. The archetype score corresponds to the standard scaled archetypal expression profiles (Methods). Heatmaps are individually column normalized for visualization. **c**, Fraction of proteins in top archetypal features for each archetype. Features in each archetype were selected in the "top" if they had an archetype score of greater than 2. For these features, we performed a one-sided hypergeometric test to determine if proteins were over-represented in this feature set relative to the global distribution of feature types. Archetypes with over-representation of proteins (one-sided hypergeometric test, BH-adjusted P<0.05) are denoted.



Extended Data Fig. 9 Visualization of archetypes in totalVI-intersect model of SLN-all. a, UMAP plots of SLN-all cells colored by latent dimension value. **b**, totalVI protein expression for CD24 and CD93 proteins as a function of distance to archetype 16. **c**, totalVI denoised expression for *Isg20* and *Ifit3* genes as a function of distance to archetype 7. Archetype is colored in blue, all other cells in grey.



Extended Data Fig. 10 totalVI identifies correlated modules of RNA and proteins that are associated with the maturation of transitional B cells. a, UMAP of the totalVI latent space colored by totalVI RNA expression of *Rag1*. **b**, totalVI RNA expression of *Rag1* as a function of 1 - Z_{16} (the totalVI latent dimension associated with transitional B cells). **c**, totalVI Spearman correlations in mature B cells between the same RNA and proteins as in Fig. 5h. Features were hierarchically clustered within mature B cells. **d**, Histogram of Spearman correlations between each feature in (**a**) and 1 - Z_{16} (n = 2,735 cells).

Chapter 3

Joint analysis of transcriptome and proteome measurements in single cells with totalVI: a practical guide

Zoë Steier, Annie Maslan, Aaron Streets†

† Corresponding author

Adapted from work submitted for publication as: Steier, Z., Maslan, A., Streets, A. "Joint analysis of transcriptome and proteome measurements in single cells with totalVI" in Single Cell 'Omics of Neuronal Cells, edited by J.V. Sweedler, J.H. Eberwine, S.E. Fraser, Neuromethods Series, Springer Nature, 2021.

Summary
The recent development of single-cell multi-omic measurement techniques necessitates analysis strategies to combine these datasets into a single view of a cell. In the previous chapter, I presented totalVI, a computational framework for joint analysis of paired RNA and protein data. Despite the availability of commercial products for performing multi-omics experiments and open-source software packages like totalVI for performing analysis, there are many details and practical challenges that scientists must overcome in order to implement published methods in real-world settings across different biological contexts and experimental designs. In this chapter, I provide a guide for fellow researchers on how single-cell multi-omics analysis of RNA and proteins can be performed in practice. I first present an overview of the experimental and computational pipelines for single-cell analysis of paired RNA and protein measurements. I then describe the practical steps necessary to complete these pipelines from collecting paired RNA and protein data from single cells to preprocessing and filtering the sequencing data, running the totalVI model, and conducting downstream analysis. I also provide notes on common pitfalls and offer recommendations so that joint analysis of RNA and proteins can be applied widely to other biological systems. This joint analysis of single-cell multi-omic data not only provides a richer definition of cell state, but also has the potential to elucidate the dynamics and mechanisms of cellular processes by characterizing the relationship between molecular layers within a cell.

Key words: single-cell analysis, transcriptomics, proteomics, multi-omics, CITE-seq, deep generative model, scRNA-seq, human cell atlas

Introduction

Single-cell measurements can reveal meaningful biological heterogeneity that is otherwise obscured by bulk analysis. Whole transcriptome analysis with single-cell RNA sequencing (scRNA-seq), for example, has been used to generate comprehensive maps of cell types and transcriptional states in both healthy and diseased primary tissue (Regev 2017). New technologies to measure multiple components of a single cell are now uncovering not only the variability within a given molecular layer, such as the transcriptome, but also the variability in the associations across layers, including the genome, proteome, and the epigenome. Such multimodal measurements are distinguished in their ability to make simultaneous orthogonal measurements on the same single cell. There are many emerging multimodal measurements that make use of tools such as imaging, sequencing, and other molecular assays (Chen 2015, Paul 2015, Wilson 2015, Chen 2020). The "multi-omic" methods that we discuss here probe large scale and genome-wide features, typically by converting the molecular components into a sequenceable readout (Figure 1a). As these single-cell multi-omic measurements mature, there is a growing need for analysis strategies to combine these datasets into a single view of a cell. Here we describe a computational pipeline for analyzing paired protein and RNA data. The approach to data analysis described here could, with modifications, be extended to other multi-omic measurements.

While high-throughput single-cell analysis technologies such as flow cytometry and scRNA-seq are independently powerful for identifying cell types and characterizing cellular heterogeneity, multi-omic analysis of single cells provides a more comprehensive view of cell state and of the relationship between molecular layers within a cell. For example, biologists have commonly relied upon an extensive literature of cell surface markers to define cell types and to sort cells for experiments, but recent cell atlas projects have amassed vast quantities of scRNA-seq data from

which cells can be defined by their unique marker genes or gene signatures. The transcriptome offers a broader view of the cell than a limited protein panel, however, scRNA-seq has limited detection efficiency of genes expressed at low levels. Combining the breadth of the transcriptome with the depth and stability of the proteome creates a more complete view of a cell's state and allows the dynamics between the two modalities to be mapped. Beyond paired transcriptome and proteome measurements, other approaches for assessing multiple modalities in parallel have been developed. Many methods jointly measure the transcriptome and epigenome, such as with an epigenomic measurement of DNA methylation, accessibility, or genome organization (Clark 2018, Cao 2018, Chen 2019, Rooijers 2019).

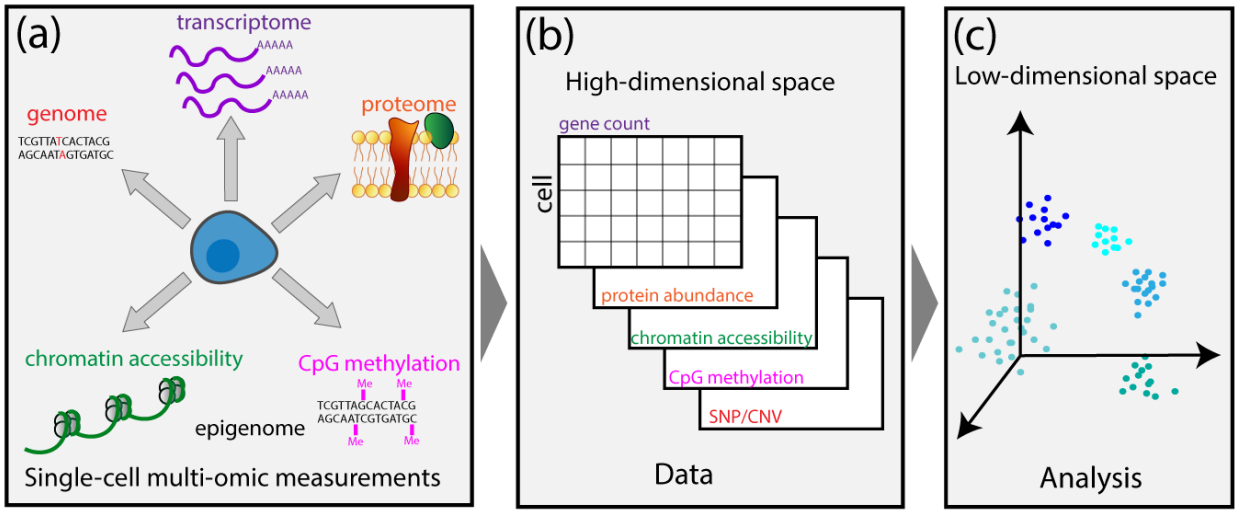


Figure 1: Single-cell multi-omics overview. a) Single-cell multi-omic measurements simultaneously profile multiple dimensions of the cell. Here, the major -omes currently used in multi-omic technologies are described: genome, epigenome, transcriptome, and proteome. **b)** For each -ome, a cell by measurement matrix is produced. **c)** These separate views can be combined to represent a cell in low dimensional space.

Multi-omic approaches not only provide a more comprehensive view of cell state, but also have the potential to illuminate the dynamics and mechanisms of cellular processes. For example, using joint measurements of chromatin accessibility and gene expression in single cells, Cao *et al.* predict gene expression from changes in chromatin accessibility at promoters and linked distal sites (Cao 2018). Similarly, using paired mRNA and protein data, Gorin *et al.* model protein translation kinetics to measure "protein velocity and acceleration" (Gorin 2020). In particular, instances where mRNA and protein measurements are not correlated suggest an active layer of post-translational regulation. Thus, multi-omic measurements can elucidate regulatory mechanisms and outcomes, whether that outcome is RNA expression, protein expression, or some other phenotype detected with imaging.

A major hurdle in working with these new types of multi-omic data sets is combining distinct data types into a single representation of a cell (Figure 1b, 1c). Each measurement presents unique technical biases, such as technical noise, limited sensitivity, background, and sparsity. Distinguishing technical noise from meaningful biological variability across omic measurements with unique noise profiles therefore complicates analysis. Processing and normalizing different datasets and combining them into a single view of a cell presents a challenge, and much of the

analysis to date has looked at modalities independently. However, methods such as MOFA, MOFA+, LIGER, and Seurat V3 have been developed to integrate multiple modalities (Argelaguet 2018, Argelaguet 2020, Welch 2019, Stuart 2019). Multi-omics factor analysis (MOFA) is a multiview matrix factorization method that identifies factors that explain variation across multi-omic datasets and has been applied to paired data sets profiling mRNA, DNA methylation, and somatic mutations or chromatin accessibility (Argelaguet 2018, Argelaguet 2019). Unlike MOFA, MOFA+ accounts for additional information about the structure between cells, such as batch and experimental conditions (Argelaguet 2020). LIGER uses an integrative non-negative matrix factorization (iNMF) to create a low-dimensional space in which a cell is defined by dataset-specific factors and a set of shared factors (Welch 2019). Seurat V3 identifies pairwise correspondences, or "anchors," between single cells across datasets to transform the datasets into a shared space (Stuart 2019).

In this chapter, we focus on joint analysis of surface protein abundance and gene expression in single cells, a rapidly growing type of multi-omic data. Recently, CITE-seq and REAP-seq have demonstrated the ability to probe these two modalities simultaneously in single cells at high throughput with whole transcriptome profiling and the measurement of an expanded panel of proteins made possible by oligonucleotide-barcoded antibodies (Stoeckius 2017, Peterson 2017).

We designed totalVI as a tool for conducting a joint analysis of RNA and protein data that addresses these assorted challenges of multi-omics analysis (Gayoso 2021). totalVI is a deep generative model that combines RNA and protein information into a single representation and leverages all available information (the gene and protein expression from each cell) to inform downstream analysis. Unlike previous strategies that analyze each modality independently and subsequently attempt to draw connections between the two, the joint analysis of totalVI combines this information to form a single, consistent view of cell state. Because of the direct connection between RNA and protein molecules through biology's central dogma, a single model that pools RNA and protein information is consistent with biological knowledge that these paired measurements were derived from molecules produced in the same single cell.

In its joint analysis, totalVI addresses the distinct sources of noise and technical bias in the RNA and protein measurements. While both RNA and protein data suffer from noise in the sampling of a limited number of molecules from each cell, other technical factors are specific to each modality. The RNA data tends to have limited sensitivity and variation in sequencing depth (where the total number of molecules detected per cell is known as the library size). The protein data is less sparse than the RNA data but tends to be obscured by background that arises experimentally from a combination of unbound ambient antibodies and non-specific antibody binding. These technical factors are modeled by totalVI so that analysis is based on the underlying biological signal rather than technical biases of the methods. totalVI can also correct technical differences between batches (known as batch effect correction) or between datasets (known as dataset integration). Unlike previous methods, totalVI is capable of integrating datasets with different protein panels (i.e., some or all proteins in a dataset are missing) and imputing the expression of the missing proteins.

To model single-cell sequencing data including noise and technical artifacts, totalVI draws conceptually on previous work modeling scRNA-seq data. To model RNA, totalVI uses a similar probabilistic modeling strategy as scVI to account for the sampling noise in the measurement

(Lopez 2018). For proteins, totalVI introduced a new model that addresses the issue of protein background by distinguishing the foreground and background components of the protein signal. totalVI combines RNA and protein data into a joint, low-dimensional representation, which is rooted in the notion that genes are often expressed in co-regulated networks or modules that can contain the information describing the high-dimensional transcriptome or proteome. This low-dimensional joint representation is derived using neural networks, which provide a powerful, non-linear approach to accurately model the data. totalVI connects this model with downstream analysis tasks, providing an end-to-end framework for the analysis of RNA and protein data that both addresses the technical factors in each measurement and takes all information into account.

Combining RNA and protein information in a joint analysis could enable the identification of new cell types or cell states that were not easily identified from a single measurement. The additional information provided by combining these two modalities might enhance the ability to detect markers in either modality, which could be particularly useful in finding surface proteins that could serve as potential drug targets, surface markers to isolate a cell population of interest, or transcription factors and enzymes that might define or regulate cells with a known cell surface phenotype. totalVI can bridge the gap between these two related views to form a more comprehensive understanding of cell state. The paired measurement of RNA and protein in the same cell might also illuminate the connection between the expression levels of these molecules in different biological settings or regulatory dynamics of transcription and translation.

Conducting a joint analysis of RNA and protein data with totalVI consists of four main steps (Figure 2):

1) Collecting a single-cell dataset with paired RNA and protein measurements
2) Preprocessing and filtering the sequencing data
3) Running the totalVI model
4) Conducting downstream analysis with totalVI.

Practical guidance in completing each of these steps is discussed in detail below.

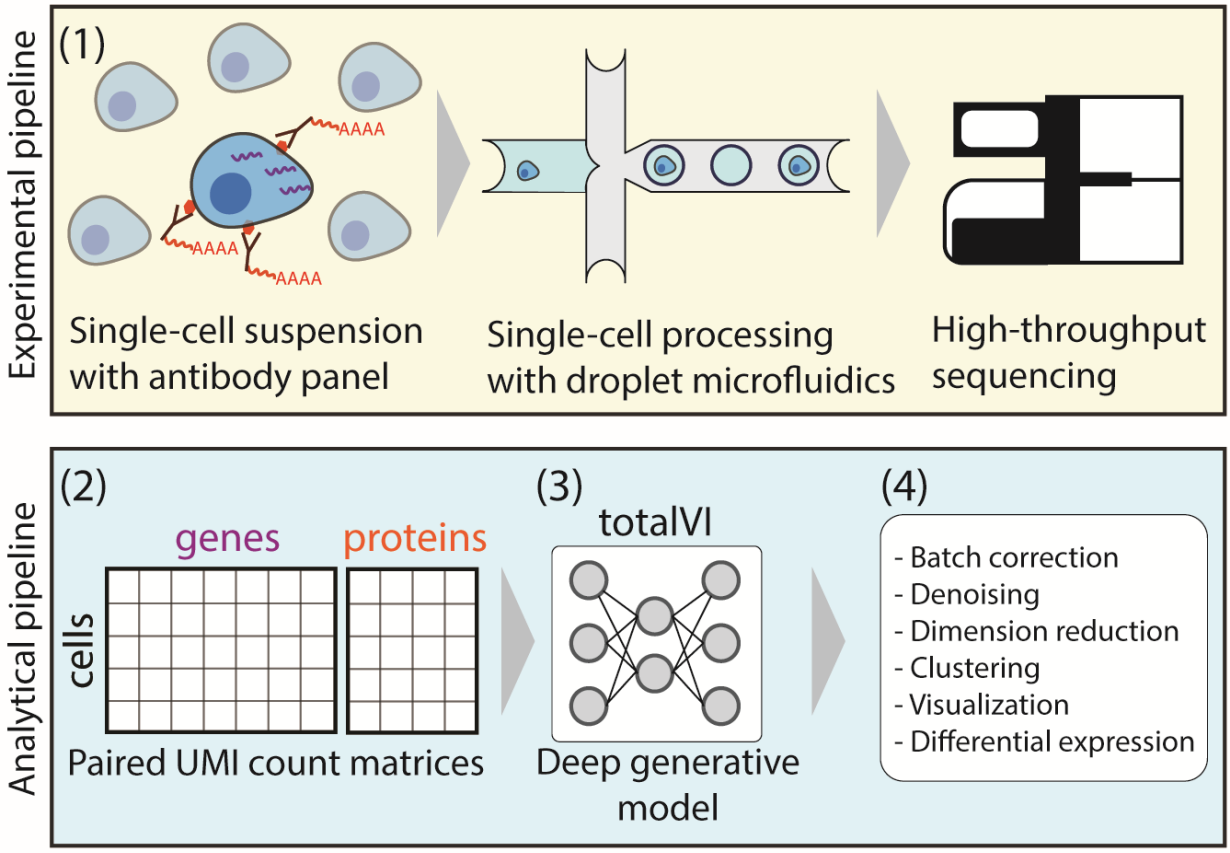


Figure 2: totalVI experimental and analytical pipeline. This schematic shows the primary steps to take a single-cell sample through joint analysis of gene expression and protein abundance with totalVI. The experimental workflow for generating a dataset containing paired RNA and surface protein abundance in single cells is illustrated in the top panel. First, a single-cell suspension is stained with a panel of barcoded antibodies. Each antibody in the panel has a unique oligonucleotide barcode that identifies the target protein and contains a poly-A tail for capture along with the polyadenylated mRNA transcripts (1). Next, this sample is processed, typically using a droplet microfluidics platform, for isolation of single cells and the reverse transcription and barcoding of captured molecules for multiplexed library preparation. After droplet processing, a single sequencing library is prepared and sequenced on a high-throughput sequencing platform. The bottom panel illustrates the totalVI analytical pipeline which requires first preprocessing and filtering the sequencing data to generate paired RNA and protein count matrices (2); Those count matrices are then used as input into the totalVI model (3). After running totalVI, downstream analysis allows for data interpretation (4).

Experimental considerations for collecting paired RNA and protein measurements in single cells

Methods for generating a single-cell dataset of paired RNA and protein measurements

Droplet-based microfluidic platforms can be used to facilitate high-throughput scRNA-seq through rapid single-cell isolation and multiplexing using oligonucleotide barcodes, which capture mRNA transcripts with a region complementary to the poly-A tail, and identify the cell-of-origin of each captured molecule. Each barcode also contains a unique molecular identifier (UMI) that enables the quantification of the original number of captured molecules after amplification (Macosko 2015, Klein 2015). Once barcoded, transcripts from thousands of single cells can be pooled together in

the same tube, reverse transcribed, amplified, and prepared for sequencing in a process known as library generation. These libraries can be further pooled across samples, sequenced, and quantified, resulting in transcript counts for each gene in each single cell.

High-throughput droplet-based scRNA-seq can be extended to measure both RNA transcripts and surface proteins from the same cell (Stoeckius 2017, Peterson 2017). Because proteins cannot be sequenced directly, these methods use antibodies specifically targeted to cell surface proteins, which are chemically conjugated to oligonucleotide barcodes that include polyadenylated tails (resembling the structure of mRNA transcripts). After staining cells with barcoded antibodies, cells are processed through droplet-generating devices, and both RNA transcripts and antibody barcodes can be captured, sequenced, and quantified in the same experiment.

Currently, this droplet-based approach is the state-of-the-art for paired RNA and protein measurements. A number of companies have commercialized oligonucleotide-barcoded antibodies that target proteins on the cell surface, including BioLegend's TotalSeq and BD's AbSeq, which are compatible with the 10x Genomics platform (Zheng 2017) and the BD Rhapsody platform, respectively. The availability of these products has made the paired measurement of RNA and proteins in single cells increasingly accessible and has enabled the application of this multi-omic technology to various cell types, organisms, and disease states (Granja 2019, Praktiknjo 2020, Kotliarov 2020, Lavaert 2020, Muench 2020).

Experimental Design

Step-by-step protocols for conducting paired RNA and protein measurements are extensively documented by the manufacturers of commercial platforms and in the original publications (Stoeckius 2017, Peterson 2017). However, there are a few aspects of experimental design that should be considered regardless of the experimental platform. These experimental decisions can affect the quality of the sequencing data collected and can have an impact on downstream analysis (see Note 1).

Cell numbers: Droplet-based microfluidic platforms can process multiple thousands of cells per reaction. For robust downstream analysis with tools like totalVI, it is recommended that an experiment contains as many cells as possible (considering sequencing costs and sample availability), but at least around 2,000 cells. Although cells can be loaded in numbers higher than those recommended by the manufacturer, caution should be taken since increasing cell numbers will increase the likelihood that two or more cells will be encapsulated in a single droplet, producing a "doublet" that should be filtered out from downstream analysis.

Cell viability: The preparation of single cell suspensions should be completed as quickly as possible to maintain cell viability. As cells die, RNA degrades more quickly than protein. Therefore, when analyzing sequencing data, low-quality RNA libraries could be an indication of poor cell viability. To preserve cell viability after tissue extraction, cells can be stained, washed, and loaded in media (including serum; see manufacturer's instructions for types of media that are incompatible with the platform in use). Commercial antibodies are formulated in buffers containing ingredients that might exacerbate cell death. To mitigate this, a buffer exchange can be performed on antibodies prior to cell staining to transfer antibodies into media or a more compatible buffer.

Cell hashing: Cell hashing is the practice of staining a sample of cells with a uniquely-barcoded antibody ("hashtag antibody") that targets a ubiquitous surface marker such that multiple samples receiving different hashtag antibodies can be pooled together in the same experiment and demultiplexed following sequencing (Stoeckius 2018). Cell hashing is employed to save costs and mitigate batch effects by processing multiple samples in the same reaction. Cell hashing also provides a way to detect doublets between samples, facilitating the overloading of microfluidic devices with higher cell numbers. It is recommended to confirm that hashtag antibodies do indeed bind to the cell type of interest, since not all cells express so-called "ubiquitous" surface proteins.

Antibody numbers: Because the available sequence space of oligonucleotide barcodes is virtually unlimited (Stoeckius 2017), the number of antibodies included per assay has thus far only been limited by the availability of barcoded antibodies. To date, experiments including hundreds (Gayoso 2021) of barcoded antibodies have reported no negative effects due to high antibody numbers, raising the potential for a future proteome-wide assay. To reduce pipetting errors and improve consistency across experiments, some companies have begun producing pre-mixed "panels" or "cocktails" of barcoded antibodies. Caution should be taken to ensure that such pre-mixed panels have similar staining conditions (see *Antibody staining* below).

Antibody titration: In some protocols, barcoded antibodies are added to the experiment in different concentrations (known as titrations). Traditionally, antibody titrations aimed to optimize the signal-to-noise ratio. Antibodies at too high a concentration might result in high levels of background due to non-specific binding or excessive unbound antibodies in the ambient solution. Alternatively, antibodies at too low a concentration might be undetectable or indistinguishable from background noise. Finding the optimal concentration for each antibody can be challenging, since this concentration is expected to differ across tissues and cell types depending on the abundance of the target protein (i.e., high concentrations are often used to detect rare/lowly expressed proteins). A recent study of barcoded antibodies explored reducing antibody titrations relative to previously published protocols with the goal of achieving sufficient signal while reducing the high cost of sequencing (Buus 2020). For instance, reducing an antibody's titration might reduce background from excess antibodies, resulting in a lower required sequencing depth to recover the same level of signal over background. While reducing antibody concentrations might reduce background and lower sequencing costs, concentrations that are too low might inhibit the distinction of protein foreground from background and could reduce the power to detect differentially expressed proteins. Moreover, adding antibodies at different concentrations makes it challenging to draw comparisons between the absolute levels of two proteins.

Antibody staining: Available protocols currently recommend uniform staining conditions for all antibodies in an experiment (e.g., add antibody panel to cell suspension and incubate for 30 minutes at 4°C). It should be noted that antibodies for some proteins (e.g., chemokine receptors) typically require staining at higher temperatures and for longer times due to protein cycling onto and off of the cell surface. Therefore, pre-mixing antibody panels and conducting a single staining step might result in poor detection of some antibodies. It is recommended to consider the ideal staining conditions of each antibody included in the assay and to potentially conduct multiple rounds of staining to allow for the binding of all antibodies of interest.

Antibody aggregation: Due to the potential of antibodies to aggregate in solution, most protocols recommend a high-speed centrifugation step prior to staining to remove antibody aggregates. It should be noted when processing sequencing results that cells with extremely high (outlier) numbers of protein counts could be caused by these antibody aggregates and should be filtered out prior to downstream analysis.

Antibody washing: Following staining, washing is an important experimental step to remove unbound antibodies and only retain antibodies on the cell surface that have specifically bound their target protein. In principle, the wash volume and number of washes could be modulated to improve the signal-to-noise ratio of barcoded antibodies. While it is recommended to use no less than the recommended wash volume or number of washes in the appropriate protocol (which could result in higher levels of protein background), attempting to reduce protein background by increasing the number of washes would come with the trade-off of reducing the overall cell number (since cells are lost in each round of centrifugation).

Sequencing depth: Sequencing depth is a major experimental consideration that can impact the quality of downstream analysis. Sequencing single-cell libraries to saturation (i.e., counting each captured molecule at least once) provides the maximum information in a single experiment, but can be cost prohibitive. Previous studies and current protocols make recommendations far below the point of saturation, but this can come at the cost of failing to detect some molecules at all (known as dropout events) or inaccurately quantifying relative molecule counts. Because RNA and protein libraries are generated separately following upstream processing steps, these libraries can be sequenced at different depths depending on the relative amounts of information provided by each molecule in a given experiment. The choice of sequencing depth typically depends on the complexity of each library (i.e., the number of unique transcripts in a cell and the total number of antibodies included in the antibody panel, see Figure 3), cost considerations, including the number of cells analyzed, and the downstream analysis goals. For example, accurate clustering and cell type identification can be achieved at relatively low sequencing depths (Heimberg 2016, Svensson 2019), but detailed downstream analysis, differential expression testing, and the detection of rare cell types or molecules might suffer at low depths. Particularly for protein analysis, sequencing at lower depth can make it more challenging to distinguish protein foreground from background. While a determination of the desired sequencing depth will depend on the cell type, proteins, and analysis of interest, it is recommended that initial experiments sequence at relatively high depth, and to reduce the sequencing depth in subsequent experiments if appropriate. For RNA libraries, a starting point is typically around 50,00-100,000 reads/cell depending on prior knowledge of low/high RNA content per cell. For protein libraries, a starting point is around 5,000-25,000 reads/cell depending on the number of antibodies in the panel (e.g., 10-100+ antibodies). RNA and protein libraries can be pooled at the desired ratio in the same sequencing run to modulate the relative sequencing depth of each library.

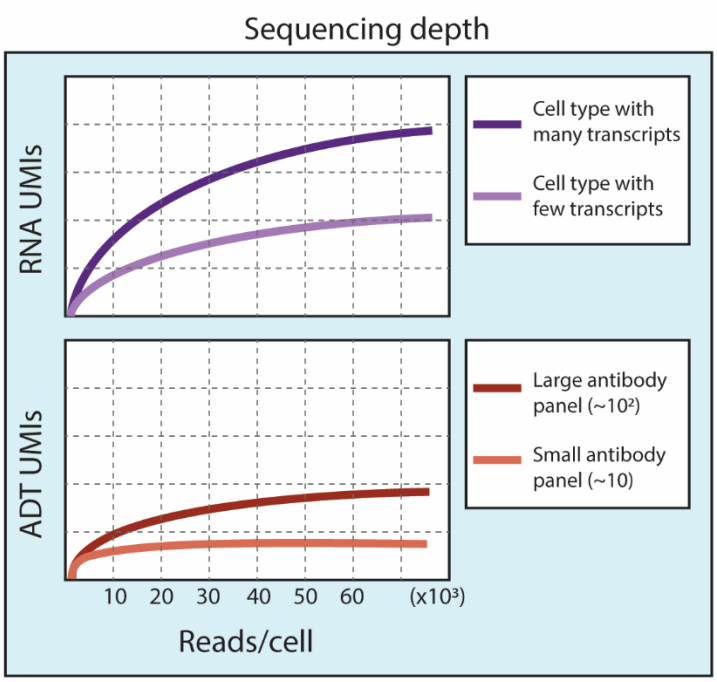


Figure 3: Sequencing depth. This schematic illustrates the relationship between sequencing depth and recovered information as measured by average uniquely captured molecules (UMIs) corresponding to RNA transcripts (top) or surface proteins (bottom). In both RNA and protein libraries, the complexity of the system will determine the sequencing depth needed before all of the available information is collected. Cells with higher RNA content, such as HEK cells, saturate at higher sequencing depths than cells with lower RNA content, such as PBMCs.

Joint analysis of paired RNA and protein measurements in single cells with totalVI

Preprocessing and filtering the data

Sequencing of RNA and protein libraries will produce reads in FASTQ format. For all analysis, FASTQ files must first be processed into a UMI count matrix, which should be filtered for quality prior to analysis with totalVI (see Note 1).

Generating count matrices: UMI count matrices can be generated from FASTQ files using tools such as Cell Ranger (Zheng 2017) or kallisto (Melsted 2019), which align sequencing reads to a reference genome (for RNA) or to a reference of predefined oligonucleotide barcode sequences (for protein). The resulting matrix will contain a row for each feature (i.e., a gene or a protein) and a column for each cell. Each entry indicates the number of captured molecules (e.g., transcripts of a given gene or antibodies binding a target protein) in a given cell.

Doublet filtering: Doublets, which are droplets that contain two or more cells rather than a single cell, can confound analysis (potentially appearing as nonexistent, hybrid cell types) and should be removed. Doublets composed of different cell types can be identified and removed by a number of existing packages (Wolock 2019, Gayoso 2018). These doublet detection packages use similar algorithms to detect doublets composed of different cell types. Caution should be taken when applying these methods to datasets containing homogeneous populations or continuous developmental trajectories since these algorithms are not expected to perform well in this setting

and could produce misleading results. If cell hashing was used, cross-sample doublets as determined by hashtag demultiplexing (Stoeckius 2018) should also be removed at this stage.

Cell filtering by quality: It is recommended to remove dead cells, low-quality cells, and doublets before analysis. For convenience, this stage of preprocessing can be performed using common single cell analysis pipelines like Scanpy (Wolf 2018) or Seurat (Stuart 2019). Typically, dead/low-quality cells can be identified by a high percentage of RNA UMI counts derived from mitochondrial genes (e.g., > 15%) and low numbers of genes/proteins detected (the number will depend on sequencing depth per experiment). It is also recommended to remove outlier cells with extremely high protein UMI counts, since this is likely caused by antibody aggregation. However, it is not recommended to remove outlier cells with high RNA UMI content in an attempt to remove potential doublets, since these high UMI counts could be driven by biological differences that are of interest for analysis (e.g., large cell size, cell cycle stage).

Feature filtering: Before running totalVI, it is recommended to conduct gene filtering. Users can decide how many genes to include in the model, but 5,000 highly variable genes is reasonable in most settings. Using fewer than 500 genes might result in lower accuracy or less informative analysis, and using greater than 10,000 genes might lead to longer run-times without any substantial improvement in performance. Hashtag antibodies and isotype control antibodies (which have no specific protein target) should also be removed prior to running totalVI.

Running the totalVI model

The totalVI software: totalVI is available as open-source, Python-based software as a part of the *scvi-tools* software package (https://scvi-tools.org/). The package includes tutorials and documentation describing how to perform the analysis pipeline described below.

Model inputs: totalVI takes as input the filtered UMI count matrix including both RNA and protein features. totalVI can also include a categorical covariate (called "batch"). Batch labels can be applied to remove batch effects, for example differences across lanes of a 10x Genomics experiment, patient samples, or sequencing runs. Batch labels can also be applied to integrate datasets derived from different experiments or labs. totalVI is capable of integrating datasets with different protein panels. As a demonstration, Chapter 2, Figure 3a-c shows UMAP plots of two CITE-seq datasets of spleen and lymph node cells with panels containing either 111 or 208 antibodies (SLN111-D1, SLN208-D2). Linear dimensionality reduction with principal components analysis (PCA, 30 components) followed by UMAP reveals batch as a major source of variation (Chapter 2, Figure 3a). To remove these batch effects, the user can choose whether to run the totalVI model on only those proteins included in both datasets (totalVI-intersect; seen in Chapter 2, Figure 3b), or to use all available protein information (totalVI-union; seen in Chapter 2, Figure 3c), which will impute the missing protein values. As can be seen in these figures, totalVI-intersect and totalVI-union both correct the batch effects between the datasets. totalVI-union can also be used to integrate an scRNA-seq dataset (containing no protein information) with a CITE-seq dataset (containing both RNA and protein). In this case, totalVI can predict protein expression in the RNA-only data using the given CITE-seq data.

Running the model: In general, it is recommended that users run the totalVI model with default hyperparameters (e.g., number of latent dimensions, learning rate, etc.). These parameters have

demonstrated accurate model fits on multiple cell types and datasets and are a good starting point for any analysis (see Note 2). The larger the size of the dataset (number of cells and number of features), the longer it will take to run the totalVI model. The model will run considerably faster on a GPU (highly recommended), which can be accessed on cloud computing platforms. After training, the model can be saved for downstream analysis at a later time.

Downstream analysis

Once trained, the totalVI model can be used for a wide variety of common analysis tasks. Here we present a common pipeline for downstream analysis with totalVI and show how totalVI can be used to learn about cells from the combination of their RNA and protein expression.

Model outputs: The totalVI model contains two different components that can be used for downstream analysis. First, totalVI contains a low-dimensional representation of cells, known as the latent space, which contains information from both the RNA and protein data while controlling for noise and technical artifacts in each data type. The low-dimensional latent space can be used to visualize and cluster cell populations. Second, totalVI contains the parameters that describe the probabilistic distributions of each observed RNA and protein measurement. These distributions account for the noise and technical artifacts in the observed data, and can be used to denoise the high-dimensional RNA and protein expression data. These distributions can also be used to test differential expression and perform other analysis tasks while accounting for the noise and technical artifacts in the RNA and protein measurements. Below, we describe how these components of totalVI can be used to conduct downstream analysis.

Dimension reduction for visualization: The totalVI model encodes the RNA and protein information from each cell in a single low-dimensional representation known as the latent space. Because the latent space includes a combination of RNA and protein information and has removed technical artifacts like batch and RNA library size, a cell's location in the latent space can be interpreted as representing the cell's underlying biological state. By applying a dimension reduction method like UMAP (McInnes 2018) to the latent space, users can visualize cells in two dimensions and interpret cell-cell similarities based on the combination of cells' RNA and protein expression profiles. An example can be seen in Chapter 2, Figure 4a, which shows a UMAP representation of the latent space from four spleen and lymph node datasets integrated with totalVI-intersect.

Clustering cells: Cells can be clustered using any clustering algorithm, such as the Leiden algorithm (Traag 2019), based on cell-cell similarity in the latent space (see Chapter 2, Figure 4a). Therefore, clusters based on the totalVI latent space take into account both the RNA and protein information from each cell. These clusters can be used to identify and annotate cell types.

Denoising features for visualization: totalVI can produce a matrix of denoised expression values for RNA and protein that remove technical effects like RNA library size, protein background, and sampling noise. The removal of protein background in denoised values can be seen in Chapter 2, Figure 2g-h. For CD20 protein, which is a marker of B cells, the distributions of raw counts are difficult to distinguish between true expression in B cells and background in T cells in the SLN111-D1 dataset (Chapter 2, Figure 2g). By denoising protein data with totalVI, the expected amount of background is effectively subtracted from the overall expected expression, resulting in a clear

separation in CD20 expression levels between B cells and T cells (Chapter 2, Figure 2h). These denoised values can be visualized to observe expression patterns across cells (see Chapter 2, Figure 5a-c). Note that denoised values can be helpful for visualization, but downstream statistical analysis should not be applied to these values directly. Instead, for downstream analyses like differential expression testing and the calculation of correlations between features, totalVI contains statistical methods that account for technical artifacts (see Note 3). In the case of totalVI-union when a dataset might be missing data from some or all proteins, imputed protein values can be visualized to observe the predicted expression patterns of these proteins (see Note 4).

Interpreting protein foreground probabilities: To accurately analyze protein data, the totalVI model learns to distinguish between protein foreground (a measurement likely derived from true biological signal) and background (a measurement likely due to ambient antibodies or non-specific antibody binding). For each protein measurement, totalVI estimates the probability that the measurement was due to foreground signal. totalVI uses these foreground probabilities to remove protein background when denoising protein data and conducting other downstream analysis tasks so that they are not biased by this technical artifact in the protein measurement.

Users might be interested in visualizing these foreground probabilities to gain intuition about which cells are "on" or "off" for a particular protein. However, it is still recommended to use denoised protein expression (rather than foreground probabilities) for interpretation of cell type expression, since denoised values account for foreground probability while preserving the dynamic range of the protein measurement. Thus, users should not be concerned if a protein that is expected to be off has a higher-than-expected foreground probability, since the magnitude of denoised expression might still be very low. If the model is uncertain for a particular protein (i.e., if many cells have intermediate foreground probabilities near 0.5), it might be an indication that the targeting antibody was poorly titrated, making it challenging for the model to distinguish foreground from background.

Differential expression testing: totalVI can be used to test differential expression for both RNA and proteins. The output of the totalVI differential expression test includes an estimate of the log-fold-change of each feature between the comparison groups (indicating the magnitude of the difference) and a Bayes factor, which can be interpreted as the significance of the difference, where a higher Bayes factor means more significant. totalVI includes a number of commonly performed differential expression tests. In a one-vs-one test, differential expression of all features is tested between two groups of cells (e.g., testing between two clusters). In a one-vs-all test, each cluster is compared with all remaining cells. An example of a one-vs-all differential expression test in B cell subsets is shown in Chapter 2, Figure 5f-g. In this case, results were filtered for significance (log Bayes factor > 0.7), ranked by median log fold change, and filtered to retain genes with non-zero UMI counts in at least 10% of the tested subset. This allowed for the identification of top differentially expressed genes and proteins in each subset. This type of test can be particularly useful for cell type annotation or identifying unique markers for a cell type. When comparing between two conditions (e.g., case vs control), the within-cluster DE test can identify differences between two labeled populations within each cluster (e.g., to identify differences between diseased and healthy T cells that are mixed within the T cell cluster).

Correlations between features: Beyond the standard set of analyses described above, users might be interested in calculating correlations between features. totalVI has no explicit information about which RNA transcript encodes which protein, so any correlation learned by the model is not predetermined by known gene-protein relationships. totalVI can calculate a denoised correlation matrix (either Pearson or Spearman) between all features that removes batch effects and technical artifacts (e.g., protein background). From the denoised matrix, users can extract a correlation of interest (e.g., between a particular gene and the protein it encodes). Correlations can also be used to cluster features into modules, which might provide insight into regulatory networks. An example of denoised feature correlations clustered into modules can be seen in Chapter 2, Figure 5h. It should be noted that correlations computed across an entire dataset might differ substantially from correlations computed within a cell type, since dataset-wide correlations might largely be driven by differences across cell types that either do or do not express a given feature.

Interpretation of the latent space: Unlike many "black-box" deep-learning models that are difficult to interpret, the latent space of totalVI has interpretable latent dimensions. Users who are interested in exploring the meaning of the latent dimensions can visualize each dimension to observe which cells have high values. A latent dimension could be interpreted as describing a particular cell type (e.g., B cells) or as describing a trend that spans cell types (e.g., cell cycle). In Chapter 2, Extended Data Figure 9, we can see an example of latent space interpretation in the plotting of the values of each of the 20 latent dimensions of the full totalVI latent space. High values in these latent dimensions highlight T cells (Z_1), B cells (Z_3), cells undergoing the cell cycle (Z_14), and transitional B cells (Z_16). Further exploration of the latent dimensions could include archetype analysis (Cutler 1994), which identifies which genes and which proteins contribute the most to each latent dimension.

Beyond standard analysis: In addition to the characterization of cell types and standard downstream tasks, totalVI can be used for a number of other types of analysis. For instance, totalVI could be used to inform aspects of experimental design such as identifying optimal antibody titrations to improve the detection of foreground and background in protein measurements. totalVI could also be used to explore the information gained per RNA or protein measurement in various analysis tasks, which could inform the desired sequencing depth of each library balanced with the cost of sequencing. Taking advantage of both RNA and protein information, totalVI could also be used to investigate basic biological questions on the relationship between RNA and protein levels, including the dynamics and regulatory processes governing transcription and translation in various cell types and experimental conditions.

Conclusions

totalVI provides a framework for end-to-end analysis of paired RNA and protein measurements in single cells. By combining RNA and protein information into a single model, totalVI creates a representation of cell state that forms a more complete description of a cell's phenotype than either modality alone. Moreover, the combination of both pieces of information about a single cell strengthens the interpretation of either modality alone. For example, the use of RNA information in the totalVI model helps to distinguish protein foreground from background, improving the accuracy of protein analysis. In addition, having protein measurements associated with single-cell

transcriptomes can help relate RNA sequencing data to prior knowledge of cell types defined by surface marker expression.

In a standard analysis with totalVI, dimensionality reduction, clustering, and visualization of denoised genes and proteins can be applied to annotate cells, potentially discovering new cell types or cell states. Through differential expression, new RNA or protein markers for these cell types could be revealed, and cells can be defined by the combination of their RNA and protein expression profiles. totalVI also provides methods to investigate correlations between features. totalVI can conduct these analyses on multiple batches or datasets through integration even if datasets did not measure identical protein panels. The totalVI model accounts for technical artifacts in each measurement, like protein background, batch effects, and sampling noise, improving the accuracy of downstream analysis.

totalVI might be extended to perform additional tasks using RNA and protein data. One potential extension could perform an automated annotation of cell types based on both RNA and protein information. In the challenging problem of learning gene regulatory networks, totalVI could provide a connection between cell surface receptors receiving signals that are transduced to produce transcriptional changes, or connecting transcriptional events with changes in surface protein expression. totalVI could eventually serve as a useful tool to query and integrate data with various published cell atlases, which contain a combination of RNA and protein information with different protein panels across tissues.

The totalVI model was designed specifically to handle the nuances of analysis of paired transcriptome and surface protein data. In the future, totalVI could be applied to additional measurements and data types. For example, as more barcoded antibodies become available, totalVI can easily scale to include the analysis of larger protein panels. It is also expected that totalVI will readily apply to the analysis of intracellular proteins measured with barcoded antibodies when these methods become available. Beyond RNA and protein measurements, the flexible framework of the totalVI model could be extended to include additional multi-omics measurements such as chromatin accessibility. For each additional modality included, the specific technical artifacts and nuances of the modality must be considered in order to properly model and address them.

Notes

1) Experimental and preprocessing choices affect downstream analysis. In particular, for data that appears low quality, the most common explanation is experimental. For RNA data that appears low quality, a suggested place to begin troubleshooting is cell viability. In the cell processing stage, even cells that are negative for markers of cell death (e.g., propidium iodide or annexin V) might have begun degrading RNA due to stress or death, so methods can be applied to preserve cell viability (as described above). Poor cell viability could also explain the capture of fewer cells than expected, since dying cells with degraded RNA often fail to meet the minimum UMI-count or gene-count thresholds for high-quality cells. For proteins that are not detected or are detected at lower-than-expected levels, it should be considered whether the antibody titration and sequencing depth are high enough to detect a signal, or if this antibody requires different staining conditions. After the data has been collected, skipping recommended preprocessing steps can result in a lower-quality analysis. Even though it might be challenging

to determine whether a certain filtering threshold is correct for a given dataset, users should not be alarmed if a few low-quality cells, dead cells, or doublets are missed by the filters. Often, after running totalVI, these cells will form separate clusters that can be filtered out at a later stage, or filtering thresholds can be updated and the model can be run again on more stringently filtered data.

2) It is highly recommended to start by running totalVI with default parameters. Advanced users might be interested in conducting parameter tuning for a particular dataset to improve performance. However, in our experience, running totalVI with default parameters provides sufficient performance across a variety of datasets that include different numbers of genes, proteins, and cells, different sequencing depths, heterogeneous cell types or more homogeneous continuous development, and different types and severity of batch effects.

3) Caution should be taken with downstream analysis on denoised values. Denoised values can be extremely helpful for visualizing RNA and protein expression after the removal of technical artifacts and sampling noise. However, due to the denoising process, denoised values might contain bias caused by spurious relationships between features. Therefore, it is not recommended for users to attempt to calculate correlations between these denoised values or to use these denoised values in a separate differential expression test. totalVI has statistical methods that control for nuisance variation while avoiding potential denoising-induced bias, which should be used for analysis testing the relationship between features or differential expression.

4) Caution should be taken when drawing conclusions from imputed protein values (e.g., when observing predicted protein expression from a scRNA-seq dataset that has been integrated with a CITE-seq dataset). These predictions can be useful for making hypotheses to be further validated experimentally and have been shown to be highly accurate in some settings (Gayoso 2021), but are expected to be less informative for cells that don't have high overlap with data where the observed proteins are present.

Acknowledgements
Research reported in this chapter was supported by the National Institute of General Medical Sciences of the National Institutes Health under award number R35GM124916. Zoë Steier was supported by the National Science Foundation Graduate Research Fellowship and the Siebel Scholarship. Annie Maslan was supported by the National Science Foundation Graduate Research Fellowship. Aaron Streets is a Chan Zuckerberg Biohub investigator. The authors would like to acknowledge Nir Yosef and Adam Gayoso for co-development of totalVI and support with the writing of this chapter.

References

Argelaguet, R. *et al.* MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, (2018).

Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

Buus, T. B., *et al.* Improving oligo-conjugated antibody signal in multimodal single-cell analysis. *bioRxiv*. doi: https://doi.org/10.1101/2020.06.15.153080 (2020).

Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science.* **361**, 1380–1385 (2018).

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science.* **348**, (2015).

Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).

Chen, T., Gupta, A., Zalavadia, M. & Streets, A. μCB-seq: Microfluidic cell barcoding and sequencing for high-resolution imaging and sequencing of single cells. *bioRxiv.* doi:10.1101/2020.02.18.954974 (2020).

Clark, S. J. *et al.* ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells e. *Nat. Commun.* **9**, 1–9 (2018).

Cutler, A., & Breiman, L., Archetypal analysis. *Technometrics*, 36(4), 338-347 (1994)

Gayoso, A. & Shor, J. *DoubletDetection* 2018. http://doi.org/10.5281/zenodo.2678042.

Gayoso, A., Steier, Z., Lopez, R. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat Methods* **18**, 272–282 (2021).

Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).

Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology* **37**, 1458–1465 (2019).

Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems* **2**, 239–250 (2016).

Klein, A. M., *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

Kotliarov, Y. *et al.* Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, **26**, 618–629 (2020).

Lavaert, M. *et al*. Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes. *Immunity* **52**, 1088-1104.e6 (2020).

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018).

Macosko, E. Z., *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).

McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. http://arxiv.org/abs/1802.03426 (2018).

Melsted, P., Booeshaghi, A.S., Liu, L. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* (2021).

Muench, D.E., Olsson, A., Ferchen, K. *et al.* Mouse models of neutropenia reveal progenitor-stage-specific defects. *Nature* **582**, 109–114 (2020).

Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015).

Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).

Praktiknjo, S. D. *et al.* Tracing tumorigenesis in a solid tumor model at single-cell resolution. *Nature Communications* **11**, 991 (2020).

Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

Rooijers, K. *et al.* Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nat. Biotechnol.* 1 (2019) doi:10.1038/s41587-019-0150-y.

Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).

Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

Svensson, V., Beltrame, E. d. V. & Pachter, L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv*. https://doi.org/10.1101/762773 (2019).

Traag, V., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).

Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

Wilson, N. K. *et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16**, 712–724 (2015).

Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, **19**, 15 (2018).

Wolock, S. L., Lopez, R., and Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, **8**, 281-291.e9 (2019).

Zheng, G. X. Y., *et al*. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, **8**, 14049 (2017).

Chapter 4

Single-cell multi-omic analysis of thymocyte development reveals NFAT as a critical driver of the CD4/CD8 lineage commitment

Zoë Steier, Laura L. McIntyre Lydia K. Lutes, Ellen A. Robey†, Aaron Streets†, and Nir Yosef†

† Corresponding authors

This work is unpublished.

Abstract
CD4 and CD8 T cells play a critical role in the mammalian immune system, and understanding their fate decisions during development has broad clinical implications relevant to autoimmune diseases such as type 1 diabetes and to the production of cancer immunotherapies. While the development of CD4 and CD8 T cells within the thymus from the CD4+CD8+ stage has been widely studied as a classic model of a lineage determination, the developmental trajectory from immature thymocytes to mature T cells and the mechanism of lineage commitment remain unclear. To deconstruct this developmental process, we apply CITE-seq to simultaneously measure the transcriptome and over 100 surface proteins in thymocytes from wild-type and lineage-restricted mice. Using totalVI, we jointly analyze the paired measurements to build a comprehensive timeline of RNA and protein expression in the CD4 and CD8 lineages. Using lineage-restricted samples, we identify early differences that implicate the calcineurin-NFAT branch of the T cell receptor signaling pathway as a putative driver of lineage commitment. Employing drug perturbations in a neonatal thymic slice system, we validate the requirement of calcium signaling through NFAT for CD4, but not CD8, lineage commitment and shed light on the CD4/CD8 lineage commitment mechanism.

Keywords
Single cell; multi-omics, thymus; thymocyte; T cell development; CD4 T cell; CD8 T cell; fate decision; lineage commitment; NFAT; calcium signaling; T cell receptor signaling

Introduction
The continuous differentiation and selection of CD4 and CD8 T cells within the thymus is critical for the maintenance of mammalian adaptive immunity. These two primary types of T cells, despite having different effector functions, arise from a common precursor cell that expresses both CD4 and CD8 surface proteins (double positive; DP). While the development of thymocytes from the DP stage into CD4 or CD8 T cells has been widely studied as a classic model of lineage determination between two irreversible fates, the mechanism of lineage commitment remains unclear. The transcription factors (TFs) THPOK (encoded by *Zbtb7b*) and RUNX3 are considered master regulators that enforce differences in phenotype and effector functions in the CD4 and CD8 lineages, respectively. However, it is unknown which upstream factors initiate the distinction between lineages that results in the differential expression of these master regulators, and how these changes occur throughout the developmental trajectory from immature thymocytes to mature T cells (Taniuchi, 2016; Saini et al., 2010).

Thymocyte populations have long been characterized into stages defined by surface protein markers, particularly the CD4 and CD8 coreceptors. However, the sorting and bulk analysis of these populations obscures the heterogeneity of continuous developmental changes and biases our understanding of this process towards prior knowledge of a limited set of surface proteins measurable by fluorescence-based flow cytometry. In contrast, advances in single cell RNA sequencing (scRNA-seq) technologies have enabled the unbiased observation of transcriptomic heterogeneity. scRNA-seq was recently used to construct a census of cells in the mammalian thymus (Park et al., 2020), opening the door to further investigations of the mechanisms underlying continuous changes along thymocyte development. Additional studies have used high-throughput single cell analysis techniques to identify the early precursor populations that seed the thymus (Lavaert et al., 2020), characterize the process of progenitor commitment the T lineage (Zhou et

al., 2019), and explore αβ T cell development (Chopp et al., 2020), yet the mechanism of CD4/CD8 lineage commitment remains an open question. While earlier studies have proposed multiple models of thymocyte development and CD4/CD8 lineage commitment (Singer et al., 2008), recent work on αβ T cells has not been able to directly address the accuracy of these models due to a lack of protein information and mechanistic study designs. Additionally, these questions are challenging to address in human or wild-type mouse samples because the fate of a DP precursor, while influenced by its T cell receptor (TCR), cannot be directly observed.

To investigate the process of thymocyte development and mechanism of CD4/CD8 lineage commitment, we performed CITE-seq to simultaneously measure the transcriptome and over 100 surface proteins in single cells from wild-type and lineage-restricted mouse thymi. Using totalVI (Gayoso et al., 2021), we jointly analyzed this data to build a comprehensive timeline of continuous RNA and protein expression changes in both the CD4 and CD8 lineages. We connected these observations to a rich literature based on fluorescence activated cell sorting (FACS) by clarifying intermediate developmental stages and defining these populations by both their transcript and surface protein composition. Furthermore, by comparing thymocyte development between lineage-restricted mice, we detected early differences in TCR signaling and identified NFAT as a putative driver of lineage differences. To validate NFAT as a differential driver of lineage commitment, we applied drug perturbations to an *ex vivo* culture system of neonatal thymic slices, avoiding genetic perturbations that could have unintended consequences upstream of the thymocyte stages of interest. While NFAT was necessary for CD4 lineage commitment, it was not necessary for either commitment to or maturation of the CD8 lineage. Our findings fill the gap in knowledge between TCR signaling at the cell surface and differential master regulator activation, establishing a model for how a fate decision is made from a common precursor.

Results

A joint transcriptomic and surface protein atlas of thymocyte development in wild-type and lineage-restricted mice

To study T cell development and lineage commitment, we profiled thymocytes from both wild-type (WT) and lineage-restricted mice. Thymocyte populations in wild-type (B6) mice closely resemble those in humans (Park et al., 2020), and serve as a model of T cell development in a healthy mammalian system. However, in a wild-type system, the ultimate fate of thymocytes as CD4 or CD8 T cells is not directly observable, making it challenging to investigate the process of lineage commitment in these divergent groups. To probe the mechanism of fate commitment, we profiled thymocytes from lineage-restricted mice including two types of MHCII-specific TCR-transgenics (AND, OT-II), two types of MHCI-specific TCR-transgenics (F5, OT-I), polyclonal MHCII-specific mice (B2M-/-), and polyclonal MHCI-specific mice (MHCII-/-). In lineage-restricted mice, thymocytes are expected to pass through the same stages of development as wild-type thymocytes (Figure 1A). However, unlike in wild-type thymocytes, the fate of lineage-restricted thymocytes is known even before cells phenotypically appear as CD4+ or CD8+ T cells, allowing for independent characterizations of CD4 and CD8 T cell development and lineage commitment.

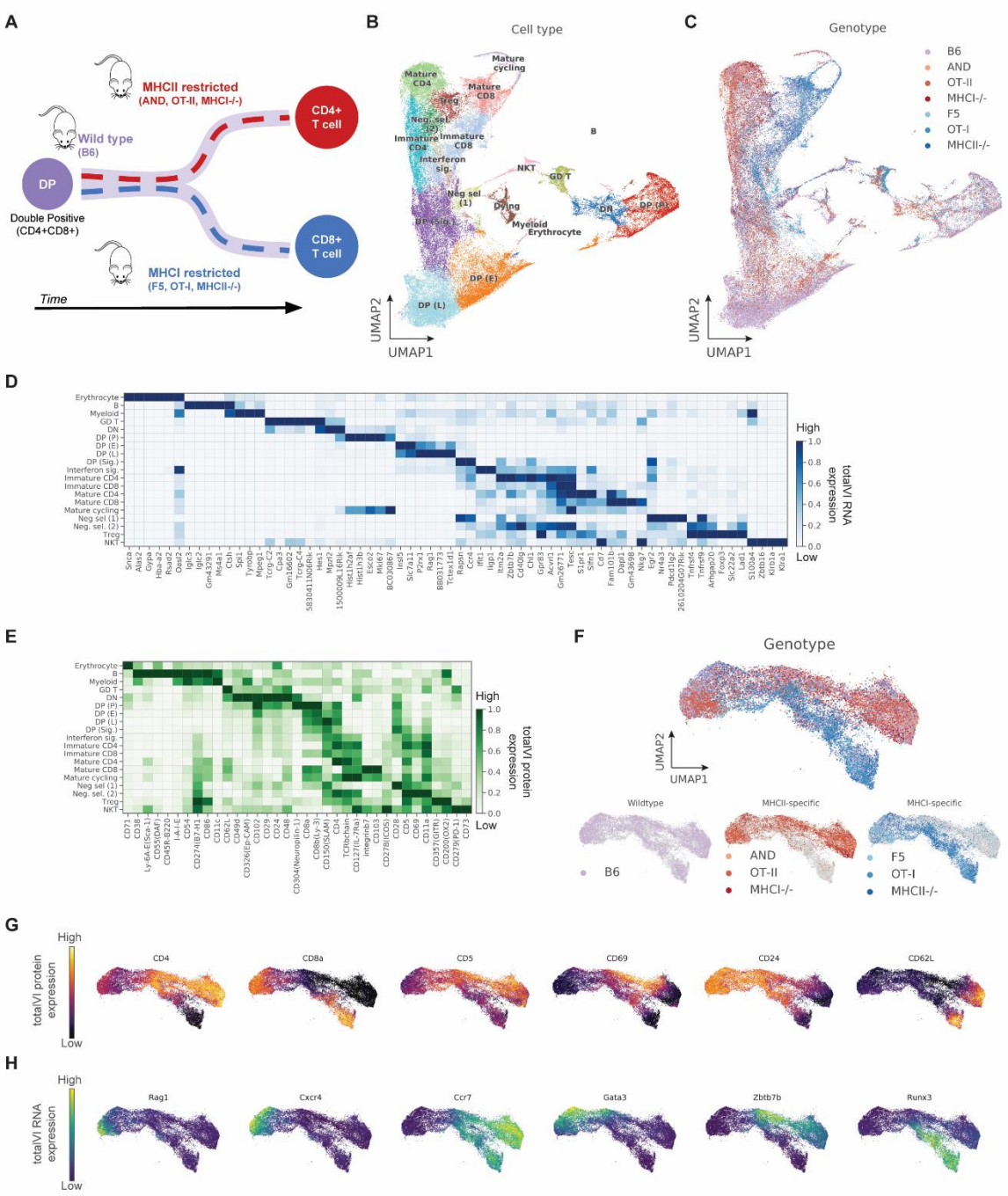


Figure 1: A joint transcriptomic and surface protein atlas of thymocyte development in wild-type and lineage-restricted mice. (A) Schematic representation of thymocyte developmental trajectories in wild-type and lineage-restricted mice used in CITE-seq experiments. **(B, C)** UMAP plots of the totalVI latent space from all thymocyte CITE-seq data labeled by cell type annotation **(B)** and mouse genotype **(C)**. **(D, E)** Heatmaps of markers derived from totalVI one-vs-all differential expression tests between cell types for RNA **(D)** and proteins **(E)**. Values are totalVI denoised expression. **(F)** UMAP plots of the totalVI latent space from positively-selected thymocytes with cells labeled by mouse genotype. **(G, H)** UMAP plots of the totalVI latent space from positively-selected thymocytes. **(G)** Cells colored by totalVI denoised expression of protein markers of lineage (CD4, CD8a), TCR signaling (CD5, CD69), and maturation (CD24, CD62L). **(H)** Cells colored by totalVI denoised expression of RNA markers of TCR recombination (*Rag1*), thymic location (chemokine receptors *Cxcr4*, *Ccr7*), and lineage regulation (transcription factors *Gata3*, *Zbtb7b*, *Runx3*).

We characterized developing thymocytes by measuring their transcriptomes and 111 surface proteins using CITE-seq (Stoeckius et al., 2017) and jointly analyzing these features using totalVI (Gayoso et al., 2021) (Methods). We sampled thymi from two biological replicates per lineage-restricted genotype and five WT biological replicates (Table S1). To enrich for thymocytes undergoing positive selection in samples from non-transgenic mice, MHC-deficient samples and three WT replicates were FACS sorted for CD5+TCRb+ (Figure S1A). We integrated CITE-seq data from all samples (72,042 cells) using totalVI, which allowed us to stratify cell types and states based on both RNA and protein information (Figure 1B). We identified all expected coarse stages of thymocyte development including early CD4-CD8- (double negative; DN) and proliferating CD4+CD8+ (double positive proliferating; DP) stages. We detected early and late stages of DP cells undergoing TCR recombination, as well as DP cells post-recombination that are downregulating *Rag* and receiving positive selection signals. In addition to immature and mature stages of CD4+ and CD8+ T cells, we observed two distinct waves of cells undergoing negative selection (Daley et al., 2013): the first appeared to emerge from the signaled DP population (lying adjacent to a cluster of dying cells), and the second from immature CD4+ T cells. *Foxp3*+ regulatory T cells appeared to cluster near mature CD4+ T cells and the second wave of negative selection. Other populations included unconventional T cells (gamma-delta T cells, NKT cells), small clusters of non-T cells (B cells, myeloid cells, and erythrocytes), a thymocyte population with high expression of interferon response genes (Xing et al., 2016), and a population of mature T cells that had returned to cycling following the cell cycle pause during thymocyte development. As expected, WT, MHCII-, and MHCI-specific samples were well-mixed in earlier developmental stages but segregated into CD4 and CD8 lineages in later-stage populations (Figure 1C).

Using totalVI, we defined cell populations not only by traditional cell type markers, but also by unbiased differential expression tests of all measured genes and proteins (Figure 1D-E, Table S3). Top differentially expressed features contained classical cell surface markers of lineage (e.g., CD4, CD8), key transcription factors (e.g., *Foxp3*, *Zbtb7b*), and markers of maturation stage (e.g., *Rag1*, *Ccr4*, *S1pr1*). Although these multi-omic definitions support the relevance of surface proteins in defining cell identities, they also reveal gradual expression changes, particularly between the DP and CD4+ and CD8+ single positive (SP) stages, that are best understood not as discrete populations, but as part of a continuous developmental process. Observation of these groups allowed us to select the populations of thymocytes receiving positive selection signals for further analysis (Methods).

We focused our analysis on developing thymocytes from the signaled DP stage through mature CD4+ and CD8+ T cells. The totalVI latent space derived from these populations captured the continuous transitions that stratified thymocytes by developmental stage and CD4/CD8 lineage (Figures 1F and S1B). Through visualization of totalVI denoised protein expression, we observed that thymocytes remained phenotypically CD4+CD8+ even after a visible branching of the lineages in the UMAP (Becht et al., 2019) representation of the totalVI latent space (Figure 1G), indicating that combining transcriptome-wide information with surface protein measurements might reveal earlier signs of lineage commitment than could have been observed from FACS-sorted populations. We also observed protein markers of TCR signaling (CD5, CD69) and maturation stage (CD24, CD62L), as well as RNA markers of TCR recombination (*Rag1)*, cell location within the thymus (chemokine receptors *Cxcr4*, *Ccr7*), and lineage regulation (transcription factors *Gata3*, *Zbtb7b*, *Runx3*) (Figure 1H). Characterization of thymocytes by their

expression of known markers relates thymocytes in our dataset to previously well-studied populations and establishes a baseline reference point for understanding surface expression and transcriptional changes over the course of thymocyte development.

Pseudotime inference captures continuous maturation trajectory and clarifies intermediate thymocyte stages

To comprehensively characterize continuous changes over the course of thymocyte development from the DP stage to the CD4 or CD8 SP stages, we used Slingshot (Saelens et al., 2019; Street et al., 2018) to perform pseudotime inference (Figures 2A and S2A; Methods). Because cell-cell similarities in the reduced dimension space are based on both RNA and protein information, both the transcriptomic and surface protein state of each cell, which are dynamic over the course of development, contribute to its placement in pseudotime. In analysis of the CD4 and CD8 lineages, we used cells from lineage-restricted mice that followed the lineage expected based on their genotype restriction.

We confirmed that the pseudotime ordering determined by Slingshot accurately captured a series of known expression events in thymocyte development (Hogquist et al., 2015) at the RNA and protein levels (Figure 2B). The expression patterns of these known features recapitulated the order of maturation events reported in prior studies (e.g., early downregulation of *Rag1* and *Rag2*, gradual downregulation of *Ccr9* and *Cd24a*/CD24, transient expression of *Cd69*/CD69, and late upregulation of *Klf2*, *S1pr1*, and *Sell*/CD62L). Many of these features associated with thymocyte maturation followed similar expression timing in the CD4 and CD8 lineages, although there were visible differences between lineages in expression levels of coreceptor molecules, master regulator transcription factors, and molecules associated with the TCR response. Because the pseudotemporal pattern of expression events that are known to correspond to maturation are well-calibrated between the CD4 and CD8 lineages at both the RNA and protein levels (Figure 2C), pseudotime values can enable comparisons between lineages of cells at comparable developmental stages. By testing for differential expression between cells grouped by pseudotime (Methods), we created a comprehensive timeline of RNA and protein expression changes in both lineages over the course of thymocyte development (Figure 2D).

We next sought to use pseudotime information to clarify the intermediate stages of thymocyte development. Thymocyte populations have been commonly defined by surface protein expression of the CD4 and CD8 coreceptors, although there is not a consensus on which intermediate populations exist in each lineage and in which order they occur (Singer et al, 2008; Germain, 2002; Bosselut, 2004; Saini et al., 2010). To clarify these intermediate populations between the DP and SP stages, we performed in silico flow cytometry analysis on totalVI denoised expression of CD4 and CD8 (Figures 2E and S2B), which showed substantial resemblance to fluorescence-based flow cytometry measurements. Thymocytes could be gated into populations based on a contour plot of cell density. Observing CD4 and CD8 expression by lineage, we found that MHCII-specific cells appeared to progress continuously in pseudotime from DP to CD4+CD8low to CD4+CD8-, while MHCI-specific cells appeared to progress from DP to the CD4+CD8low gate before reversing course to reach the CD4+CD8- gate later in pseudotime. In further detail, separating these cells into eight bins uniformly spaced in pseudotime (Figure 2F) revealed two distinct features of

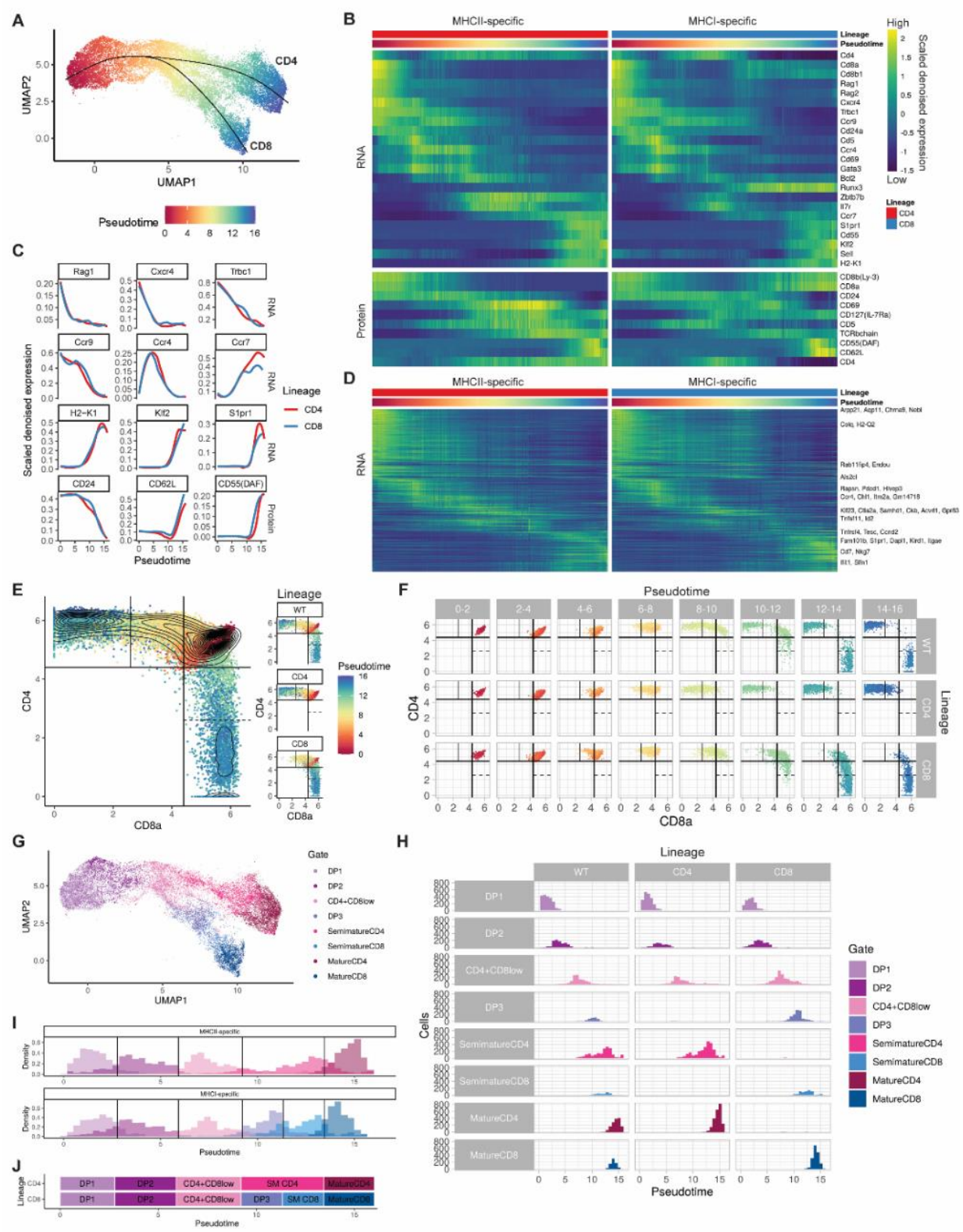


Figure 2: Pseudotime inference captures continuous maturation trajectory and clarifies intermediate thymocyte stages. (caption on following page)

Figure 2: Pseudotime inference captures continuous maturation trajectory and clarifies intermediate thymocyte stages. (A). UMAP plot of the totalVI latent space from positively-selected thymocytes with cells colored by Slingshot pseudotime and smoothed curves representing the CD4 and CD8 lineages. **(B)** Heatmap of RNA (top) and protein (bottom) markers of thymocyte development over pseudotime in the CD4 and CD8 lineages. Features are colored by totalVI denoised expression, scaled per row, and sorted by peak expression in the CD4 lineage. Pseudotime axis is the same as in (A). **(C)** Expression of features in the CD4 and CD8 lineages that vary over pseudotime. Features are totalVI denoised expression values scaled per feature and smoothed by loess curves. **(D)** Heatmap of all RNA differentially expressed over pseudotime in any lineage. Features are scaled and ordered as in (B). Labeled genes are highly differentially expressed over time (Methods). **(E)** In silico flow cytometry plots of log(totalVI denoised expression) of CD8a and CD4 from positively-selected thymocytes (left) and the same cells separated by lineage (right). Cells are colored by pseudotime. **(F)** In silico flow cytometry plot of data as in (E) separated by lineage and pseudotime. **(G)** UMAP plot of the totalVI latent space from positively-selected thymocytes with cells colored by gate. Cells were computationally grouped into eight gates using CD4, CD8a, CD69, CD127(IL-7Ra), and TCRbchain. **(H)** Histograms of cells separated by lineage and gate with cells colored by gate as in (G). **(I)** Stacked histograms of gated populations in MHCII-specific (top) and MHCI-specific (bottom) thymocytes, with thresholds classifying gated populations over pseudotime (Methods). **(J)** Schematic timeline aligns pseudotime with gated populations, with population timing determined as in (I).

MHCI-specific development. First, there exists a developmental stage (pseudotime 6-8) at which nearly all MHCI-specific cells fall in the CD4+CD8low gate, implying that this is a required intermediate stage for all CD8 T cells, not just for those with high self-reactivity. The high numbers of MHCI-specific cells in the CD4+CD8low gate underscore the fact that WT cells within the CD4+CD8low gate cannot be assumed to be committed to the CD4 lineage. Second, at subsequent times (pseudotime 8-12), MHCI-specific thymocytes pass through a DP phase on their way towards the CD4lowCD8+ and CD4-CD8+ gates while the MHCII-specific lineage does not contain late-time DP cells. While a population of later-time DP cells has been previously described (Saini et al., 2010) it is not commonly used (Park et al., 2020; Chopp et al., 2020), resulting in a missing stage of CD8 T cell development as well as a DP gate contaminated by later-time CD8 lineage cells.

For the sake of experimental isolation of intermediate thymocyte stages across time and lineage, we used pseudotime information along with surface protein data to identify a minimal set of robust surface protein markers. Four stages in time could be separated by in silico flow cytometry gating on CD69 and CD127(IL-7Ra), in which thymocytes begin with low expression of both markers, first upregulate CD69, later upregulate CD127, and finally downregulate CD69 (Figure S2C). The addition of CD4 and CD8 as markers allow for the separation of lineages at later times. Finally, the later-time DP population that is prominent in the CD8 lineage can be distinguished from earlier DP cells by high expression of TCRb (Saini et al., 2010) in addition to expressing both CD69 and CD127 (Figure S2D-E). We refer to this later-time DP population as DP3 to distinguish it from the earlier DP1 (CD69-, CD127-) and DP2 (CD69+, CD127-) populations. In combination, a gating scheme based on these five surface proteins could identify eight populations (Figure 2G-H) that allow FACS to approximate the binning of thymocytes by pseudotime and lineage (Figure 2I-J). These findings allowed us to specify an updated model of intermediate thymocyte populations in both the CD4 and CD8 lineages (Figure S2F). Fluorescence-based flow cytometry successfully replicated CITE-seq-derived gates on these surface markers, enabling the isolation of the eight described populations (Figure S2G) and supporting the presence of intermediate stages as specified in Figure S2F.

Paired measurements of RNA and protein reveal the timing of major events in CD4/CD8 lineage commitment

Previous studies have long recognized that CD4 and CD8 coreceptor expression on the cell surface, master regulator transcription factor activation, and TCR signaling play assorted roles in lineage commitment, but the relative timing and levels of these expression and signaling events in thymocyte development remain unclear. Multiple models have been proposed to describe how these components behave over the course of thymocyte development and how they might interact to initiate and enforce commitment to the CD4 and CD8 lineages (Singer et al., 2008). These models were primarily based on evidence from FACS-sorted populations that lacked high resolution in time and quantitative comparisons in expression levels. Here, we use paired RNA and protein measurements from continuously developing thymocytes to resolve the relative timing and between-lineage expression differences for these key molecular events.

The expression levels of CD4 and CD8 coreceptors play an important role not only in defining the lineage of mature T cells, but also in transmitting the TCR signals that are necessary for thymocyte development. We observed that coreceptor expression followed an expected pattern by which RNA expression changed first, followed by a corresponding change in protein expression, likely explained by the lag between RNA transcription and protein translation (Figures 3A and S3A). Beginning from the DP stage with high expression of both coreceptors in both lineages, we observed small dips in expression of both coreceptors ("double dull" stage) followed by a rise in CD4 and a continued fall in CD8a in both lineages. Eventually, in the CD8 lineage, CD8 rose and CD4 expression fell, resulting in a late transient DP stage (DP3) as the cells move towards the CD8 SP phenotype. Differential expression between the CD4 and CD8 lineages indicated that a significant difference in *Cd8a* RNA expression began building from pseudotime point 6 (approximately in the early CD4+CD8low gate), followed by a corresponding difference in CD8 protein expression (Figure 3B). It was not until later, beginning with pseudotime point 9 (the point at which the lineage can first be distinguished by flow in the semimature CD4 vs DP3 gates), that a significant difference in CD4 expression emerged.

Transcription factors THPOK and RUNX3 have been shown to either activate or inhibit the expression of the respective coreceptors later in development (Singer et al., 2008), but when differences in master regulator expression first emerge remains unclear. *Gata3*, which has been shown to play an important role in activating *Zbtb7b* expression in the CD4 lineage (Wang et al., 2008), was expressed at a higher level over a longer time in the CD4 lineage relative to the CD8 lineage (Figure 3C). Expression of *Gata3* was followed by increased expression of *Zbtb7b*, and subsequently *Runx3*. Although previously described as having mutually exclusive expression and being mutually inhibitory (Vacchio and Bosselut, 2016), we observed transient, low expression of *Zbtb7b* in the CD8 lineage at the CD4+CD8low stage simultaneous to the rise in *Zbtb7b* expression in the CD4 lineage, as well as low *Runx3* expression in the CD4 lineage throughout later time (Figure 3E and S3B). Intracellular staining in WT and MHCI-specific CD4+CD8low thymocytes confirmed the presence of a small population of cells co-expressing THPOK and RUNX3 (Figure S3C). Despite this observation that both lineages have at least minimal expression of all three transcription factors, there was a clear trend in differential expression between the lineages. First *Gata3* was upregulated in the CD4 lineage, which was followed by *Zbtb7b* upregulation in the CD4 lineage, and then by *Runx3* upregulation in the CD8 lineage (Figure 3D). Intracellular staining supported the observed timing in differential expression, detecting higher GATA3 in

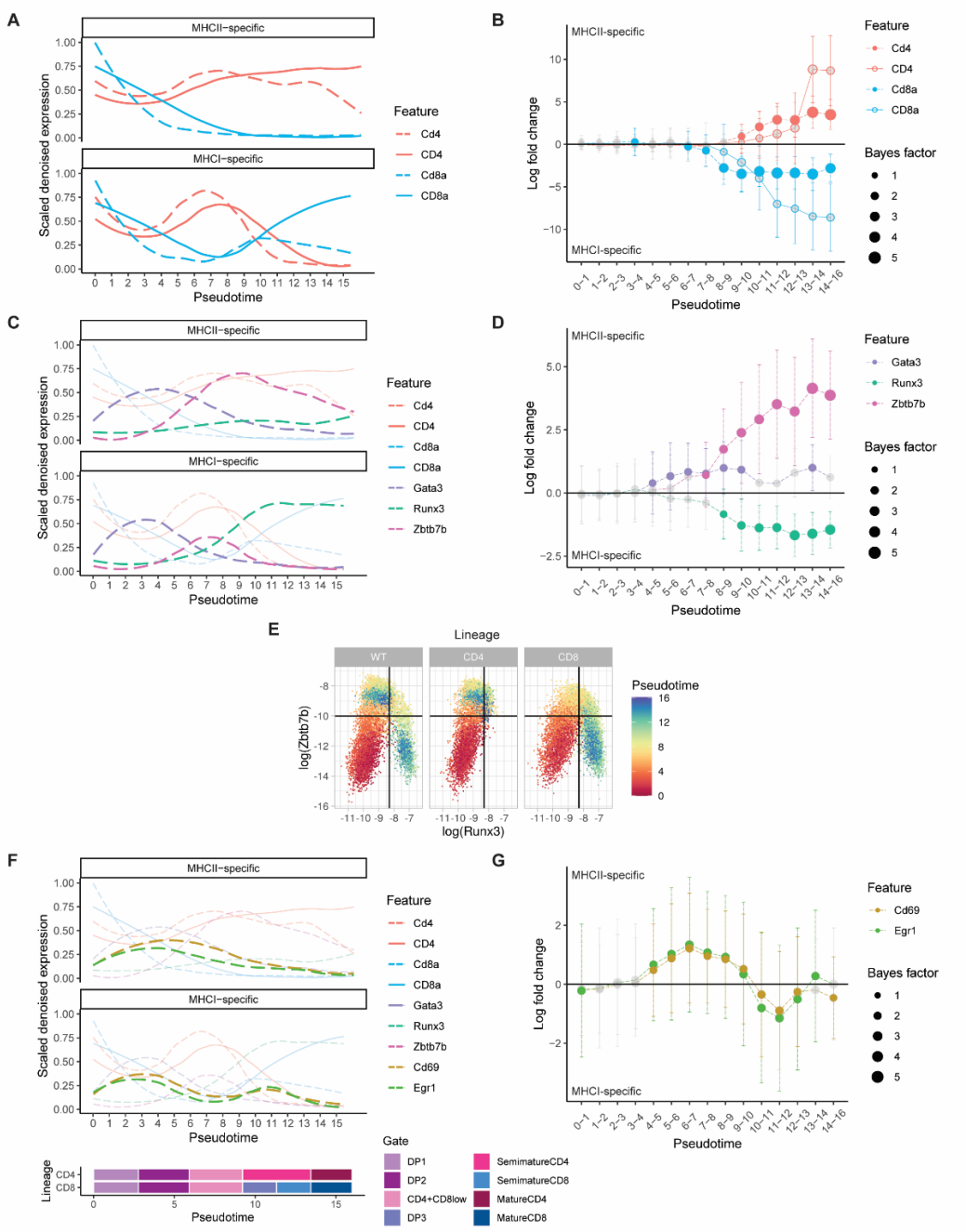


Figure 3: Paired measurements of RNA and protein reveal the timing of major events in CD4/CD8 lineage commitment. (A) Expression of co-receptor RNA (dashed) and protein (solid) over pseudotime in the CD4 (MHCII-specific) and CD8 (MHCI-specific) lineages. Features are totalVI denoised expression values scaled per feature and smoothed by loess curves. **(B)** Differential expression over pseudotime between CD4 and CD8 lineages for features in (A). Non-significant differences are gray, significant RNA results are filled circles, and significant protein results are open circles. Error bars indicate the totalVI-computed standard deviation of the median log fold change. **(C)** Expression over pseudotime as in (A), overlaying RNA expression of key transcription factors. **(D)** Differential expression over pseudotime as in (B) for features in (C). **(E)** In silico flow cytometry plots of log(totalVI denoised expression) of *Runx3* and *Zbtb7b* from positively-selected thymocytes separated by lineage and colored by pseudotime. **(F)** Expression over pseudotime as in (C), overlaying RNA expression of TCR signaling response molecules. **(G)** Differential expression over pseudotime as in (B) for features in (F). Schematic timeline aligns pseudotime with gated populations (see Figure 2J).

MHCII-specific thymocytes at the CD4+CD8low stage, higher THPOK in MHCII-specific thymocytes at both the CD4+CD8low and CD4 SP stages, and higher RUNX3 in MHCI-specific thymocytes beginning at the DP3 stage and continuing through the CD8 SP stages (Figure S3D). Considering the patterns in coreceptor expression, it appeared that CD4 was downregulated in the CD8 lineage only after *Runx3* upregulation (consistent with RUNX3 regulation of the CD4 locus (Vacchio and Bosselut, 2016)), but that CD8 expression in the CD8 lineage began upregulation prior to *Runx3* expression (Figure 3B,D). This implies that differential regulation of CD8 occurs upstream of master regulator RUNX3 expression. In addition, the later upregulation of *Runx3*/RUNX3 in the CD8 lineage relative to *Zbtb7b*/THPOK in the CD4 lineage suggests that uncommitted thymocytes might first have the opportunity to commit to the CD4 fate and only later have the opportunity to commit to the CD8 fate, contradicting the "fork in the road" model of lineage commitment.

Although it is unknown which initial factors drive the early differences in uncommitted thymocytes that lead to differential master regulator expression, previous work has pointed to TCR signaling as a potential source of difference (Germain, 2002). MHCII-specific thymocytes have been documented as having higher intensity, duration, or frequency of TCR signaling (Bosselut, 2004), but thymocytes are also known to tune their levels of TCR signaling machinery to maintain adequate signaling through positive and negative selection (Azzam et al., 1998). We found a higher and more prolonged TCR response (indicated by *Cd69 and Egr1* gene expression) in the CD4 lineage (Figure 3F), which became significantly differentially expressed simultaneous to *Gata3* and prior to *Zbtb7b* (Figure 3G), suggesting that differences in TCR signaling might play a role in driving differential master regulator activation. Unexpectedly, we also found what appeared to be a second, lower TCR response in the CD8 lineage after master regulator induction that was not present in the CD4 lineage, which might provide evidence of MHCI-specific thymocytes tuning their TCR response by increasing their sensitivity to TCR signals. The timing of these assorted events can be summarized into a temporal model of CD4 and CD8 T cell development (Figure S4).

Emergence of differences between CD4 and CD8 lineages implicate putative drivers of lineage commitment

To better understand the process of lineage commitment, we investigated how differences emerge between the CD4 and CD8 lineages. In a differential expression test between lineages, there were not substantial differences in either RNA or protein expression at the early DP stages (Figure 4A), emphasizing the phenotypic similarity in thymocytes before lineage commitment. However, differences gradually accumulated over pseudotime. To summarize differences in expression patterns over time, we clustered all genes differentially expressed at any point in pseudotime according to the lineage in which they were upregulated (Figure 4B-C). While differences that appear later in time (after master regulator expression) are often related to effector functions, differences that appear relatively early in time might contain drivers of lineage commitment. For example, MHCII-specific clusters 4 and 7 contain *Gata3*, *Cd69*, *Egr1*, and other TCR response genes, which might be early drivers of differential activation of *Zbtb7b* and *Runx3*, which themselves induce and enforce large differences. Observing the fold change in expression between lineages for all differentially expressed genes (Figure 4D-E), we found many genes following an expected pattern in becoming increasingly different over time, particularly in the CD4 lineage. In the CD8 lineage, we found some gene clusters displaying delayed expression of genes that were

previously upregulated in the CD4 lineage, implying later commitment to the CD8 lineage from an uncommitted stage.

Narrowing our focus upstream of master regulator expression, we sought to identify potential drivers of early lineage differences that result in lineage commitment. We performed transcription factor enrichment analysis with ChEA3 (Keenan et al., 2019), which identifies the transcription factors most likely to explain the expression of a set of target genes based on databases including ENCODE and ReMap ChIP-seq experiments (Dunham et al., 2012; Cheneby et al., 2020). In our study, we used genes differentially expressed between lineages in each unit of pseudotime as the target gene set (Figure 4F-G). We ranked candidate driver transcription factors based on enrichment in the three pseudotime units prior to master regulator differential expression in each lineage, and filtered out transcription factors in times at which they were not expressed in at least 5% of the population of interest (Methods). In addition to their ranking, we made note of whether each transcription factor had a known association with TCR signaling (Kandasamy et al., 2010; Methods), was known according to ChEA3 databases to be a likely regulator of *Gata3*, *Zbtb7b*, or *Runx3*, and whether this transcription factor itself was differentially expressed at the relevant pseudotime stage. We observed that multiple candidate transcription factors highly ranked in the CD4 lineage were members of pathways associated with TCR signaling (*e.g.*, *Egr2*, *Nfatc2*, *Egr1*, *Nfatc1*), and that multiple TCR response genes appeared upregulated in the CD4 lineage relative to the CD8 lineage prior to lineage branching (Figure 4H). In the CD8 lineage, multiple highly ranked transcription factors such as *Ets1* and *Tcf7* had known associations with thymocyte development (Zamisch et al., 2009), but have been previously shown to have relevance to the maturation of thymocytes in both lineages (Wang et al., 2010; Steinke et al., 2014).

In selecting candidate drivers of lineage commitment for validation studies, we prioritized transcription factors that were highly enriched in the CD4 lineage prior to master regulator activation and had known associations with TCR signaling, which has previously been suspected as an upstream source of lineage differences in uncommitted thymocytes. TCR engagement with peptide:MHC (pMHC) activates a series of signaling cascades leading to nuclear translocation and subsequent control of downstream target genes. TCR signaling occurs through three main signal transduction pathways: the calcineurin-NFAT axis, the Erk-MAP Kinase (MAPK) cascade, and the PKC-NF-kB pathway, reviewed in (Navarro and Cantrell, 2014; Hogquist and Jameson, 2014; Malissen et al., 2014; Chakraborty and Weiss, 2014). Following TCR stimulation, T cells are particularly sensitive to calcium signaling transduced through calcineurin, resulting in NFAT (NFATC1, NFATC2) translocation to the nucleus. Calcineurin and NFAT are required for positive, but not negative selection (Gao et al., 1988; Jenkins et al., 1988; Shi et al., 1989; Wang et al., 1995). The location and duration of Erk-MAPK signaling in thymocytes also determines positive versus negative selection (Daniels et al., 2006). Brief, low-intensity Erk activation downstream of TCR signaling is required for positive selection, whereas negative selection induces rapid and robust Erk signaling (Daniels et al., 2006; McNeil et al., 2005). Of note, EGR family members, *Egr1* and *Egr2*, lie downstream of the Erk-MAPK cascade (Figure 4I). The Erk-MAPK cascade has been proposed to be involved in the development of both CD4 and CD8 T cells (Sharp et al., 1997; Wilkinson and Kaye, 2001). TCR:pMHC engagement also facilitates nuclear translocation of the TF NF-kB. NF-kB is upregulated in DP thymocytes during positive selection, as well as in both CD4 SP and, to a greater extent, CD8 SP cells. However, the requirement for NF-kB in selection is unclear due to redundancy in the pathway (Hettman and Leiden, 2000; Jimi

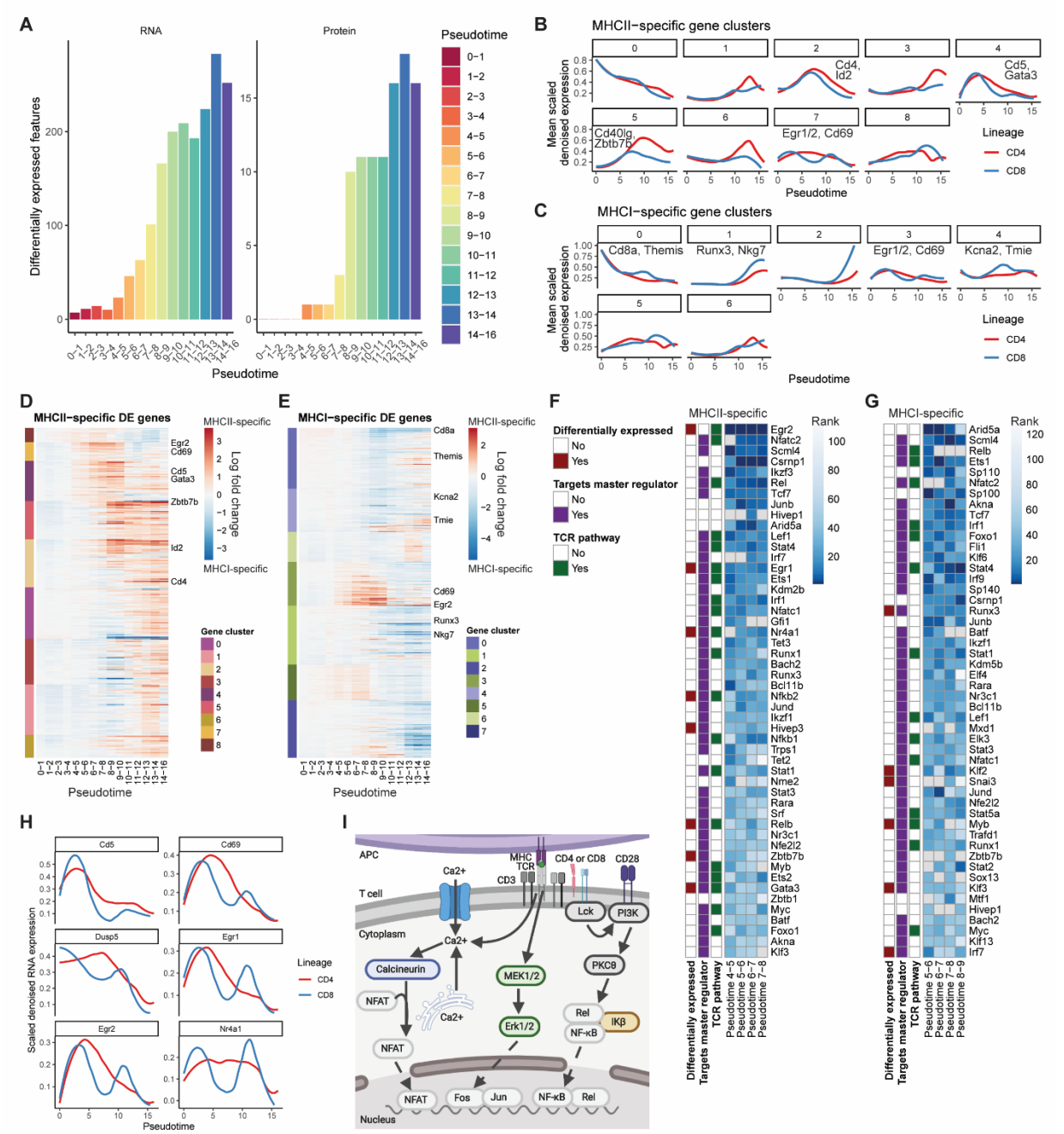


Figure 4: Emergence of differences between CD4 and CD8 lineages implicate putative drivers of lineage commitment. (caption on following page)

Figure 4: Emergence of differences between CD4 and CD8 lineages implicate putative drivers of lineage commitment. (A) Number of differentially expressed features between the CD4 and CD8 lineages across 15 pseudotime bins. **(B)** Genes upregulated in the CD4 lineage relative to the CD8 lineage scaled per gene and clustered by the Leiden algorithm according to expression in the CD4 lineage. Expression over pseudotime per cluster is displayed as the mean of scaled totalVI denoised expression per gene for genes in a cluster, smoothed by loess curves. **(C)** Same as (B), but for genes upregulated in the CD8 lineage relative to the CD4 lineage, clustered according to CD8 lineage expression. **(D)** totalVI median log fold change over pseudotime of genes upregulated in the CD4 lineage relative to the CD8 lineage. Genes are grouped by cluster in (B). Clusters are ordered by their average highest magnitude fold change. **(E)** totalVI median log fold change over pseudotime of genes downregulated in the CD4 lineage relative to the CD8 lineage (i.e., upregulated in the CD8 lineage). Genes are grouped by cluster in (C). Clusters are ordered by their average highest magnitude fold change. **(F)** Transcription factor enrichment analysis by Chea3 for CD4-lineage-specific differentially expressed genes. Transcription factors are ranked by mean enrichment in the three pseudotime bins prior to Zbtb7b differential expression (between pseudotime 4-7). Gray indicates a gene detected in less than 5% of cells in the relevant population. "Differentially expressed" indicates significant upregulation in at least one of the relevant time bins. "Targets master regulator" indicates a transcription factor that targets either Gata3, Runx3, or Zbtb7b in Chea3 databases. "TCR pathway" indicates membership in NetPath TCR Signaling Pathway or genes transcriptionally upregulated by TCR signaling, or genes with literature support for TCR pathway membership (Methods). **(G)** Same as in (F), but for the CD8 lineage, with ranking by mean enrichment in the three pseudotime bins prior to Runx3 differential expression (between pseudotime 5-8). **(H)** Expression of genes downstream of TCR signaling in the CD4 and CD8 lineages over pseudotime. totalVI denoised expression values are scaled per gene and smoothed by loess curves. **(I)** Schematic of the major branches of the TCR signaling pathway: calcineurin-NFAT (blue), Erk-Mapk (green), and NFKB (orange).

et al., 2008). Since all three TCR signal transduction pathways are involved in positive selection, it has been difficult to evaluate how TCR signaling contributes to either CD4 or CD8 lineage commitment. Based upon previously published experiments and the data presented here, we postulated that TCR signaling is not only essential for positive and negative selection, but influences CD4 versus CD8 lineage commitment. We hypothesized that thymocytes experience different timing and intensities of TCR signals resulting in preferential activation of different TCR signaling pathways, which ultimately influences commitment to the CD4 or CD8 lineage. The factors EGR and NFAT, which were both enriched for activity in the CD4 lineage, represent two of the three major branches of the TCR signaling pathway. Of note, while *Egr1* and *Egr2* themselves are differentially expressed between lineages, *Nfatc2* and *Nfatc1* are not. However, this is not surprising, since NFAT activation is regulated by calcium signaling (not observed at the transcriptional level) and could be differentially activated without being differentially expressed. Due to previous findings on the role of the Erk-MAPK cascade in both lineages, we hypothesized that the NFAT signaling branch of the TCR pathway could drive the distinction in commitment towards the CD4 or CD8 lineage.

NFAT drives commitment into the CD4 lineage via GATA3

To test our hypotheses about lineage commitment drivers in the TCR signaling pathway, we developed an *ex vivo* neonatal thymic slice culture system in which we could perturb the pathways of interest in developing thymocytes with pharmacological agents. The thymic slice system is a powerful tool to study and manipulate T cell development spatially and temporally in a three-dimensional live organ culture (Dzhagalov et al., 2012). The adult thymic slice system reliably recapitulates many temporal and phenotypic characteristics of thymocytes development reported *in vivo* (Dzhagalov et al., 2012; Melichar et al., 2013; Weist et al., 2015; Ross et al., 2014). However, because the adult thymic slice system does not reliably support the development of CD4 thymocytes, we adapted this system to track development in neonatal slices, which produce both CD4 and CD8 thymocytes (Methods).

We harvested and generated thymic slices from postnatal day 1 mice, a time point that allows us to track a synchronous wave of developing CD4 and CD8 thymocytes since T cells in mice do not develop until birth (Kernfeld et al., 2018), and cultured slices for up to 96 hours on tissue culture inserts (Figure S5A). Using flow cytometry, we quantified populations of developing thymocytes within neonatal slices (Figures 5A and S5B) based upon cell surface marker expression. The neonatal slice system supports development of both CD4+ and CD8+ T cell lineages in WT mice (Figure S5C-J). We observed a decrease in unsignaled double positive (unsig DP; CD4+CD8+CD69-) cells and an increase in semimature and mature T cell populations over the span of 96 hours (Figure S5D,H,J). We first noticed an increase in CD4+CD8lo (CD4+CD8+TCRβ+) cells after 48 hours of culture (Figure S5F); followed by an increase in CD4+ semimature (CD4+ SM; CD4+CD8-TCRβ+CD69+) cells after 48-72 hours of culture (Figure S5G), and CD4+ mature (CD4+ Mat; CD4+CD8-TCRβ+CD69-) cells after 72-96 hours (Figure S5H). Consistent with previously published results (Kurd and Robey, 2016), CD8 T cell development was slightly delayed compared to that of CD4 T cells. We observed an increase in frequency of CD4loCD8+ (CD4loCD8+TCRβ+) cells after 72 hours (Figure S5I), and mature CD8+ cells (CD8+ Mat; CD4-CD8+TCRβ+CD69-) after 72-96 hours (Figure S5J). We also observed similar developmental patterns in MHCII-specific (β2M-/-) slices, with the exception of CD4loCD8+ and CD8+ Mat T cells, and in MHCI-specific (MHCII-/-) slices, with the exception of CD4+ SM and CD4+ Mat cells (data not shown). WT, β2M-/-, and MHCII-/- mice all have a population of CD4+CD8lo cells (Figure S5K), albeit it is reduced in MHCII-/- mice. This supports the concept that all developing T cells, regardless of MHC specificity and ultimate lineage choice, first downregulate CD8, proceeding through the CD4+CD8lo stage (Singer et al., 2008). Together these observations validate that the neonatal slice system supports the development of both CD4 and CD8 lineage cells. Additionally, these data, in conjunction with previously published literature and our pseudotime analysis, support the order of CD4 and CD8 development, shown in Figure 5A, with CD4+CD8lo cells developing first, followed by the CD4 lineage (CD4+ SM then CD4+ Mat) or, alternatively, the CD8 lineage (CD4loCD8+ and CD8+ Mat).

To directly test the involvement of TCR signaling through the calcineurin-NFAT axis in lineage commitment, we inhibited calcineurin activity by adding cyclosporin A (CsA) (Liu, 1993) to neonatal slice cultures (Figure 5B). Inhibition of calcineurin-NFAT reduced early CD4+, but not CD8+ T cell populations. We observed a dose- dependent reduction in CD4+CD8lo cells in WT and MHCII-specific, but not MHCI-specific cells when CsA was added to neonatal slice cultures (Figure 5C). Addition of CsA also reduced CD4+ SM (Fig. 5D), but did not have an effect on CD4+ Mat (Figure 5E), CD4loCD8+ (Figure 5F), or CD8+ Mat (Figure 5G) populations. Notably, CsA treatment did not affect cell viability (Figure S5L), DN (Figure S5M), Unsig DP (Figure S5N), or Sig DP (Figure S5O) cell populations at low concentrations (50, 100, or 200 ng/mL). Although, we did observe a significant increase in DN cells (Figure S5P) at high concentrations of CsA (400 and 800 ng/mL), suggesting a block in b-selection which has been previously reported (Gallo et al., 2007; Urdahl et al., 1994). Together, these data suggest that NFAT is required for early CD4+ lineage commitment, but is not necessary for CD4+ maturation or CD8+ lineage commitment.

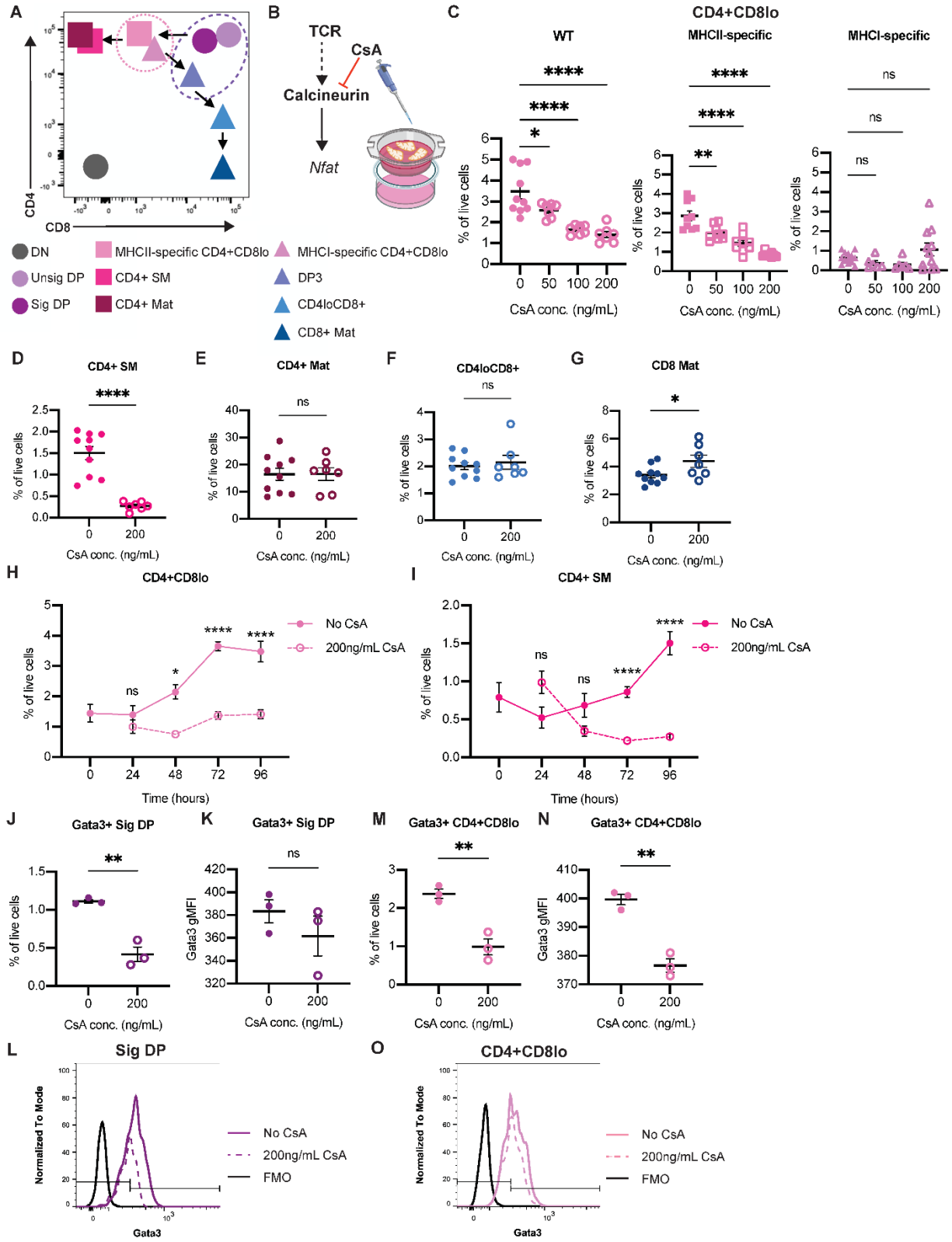


Figure 5. Inhibition of calcineurin-NFAT prevents early CD4 lineage commitment. (caption on following page)

Figure 5. Inhibition of calcineurin-NFAT prevents early CD4 lineage commitment. (A) Schematic showing cell surface markers used to identify populations of thymocytes; thymocytes were categorized into double negative (DN; CD4-CD8-), unsignaled double positive (Unsig DP; CD4+CD8+CD69-), signaled double positive (Sig DP; CD4+CD8+CD69+), MHCII-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loTCRβ+), CD4+ semimature (CD4+ SM; CD4+CD8-TCRβhiCD69+), CD4+ mature (CD4+ Mat; CD4+CD8-TCRβhiCD69-), MHCI-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loTCRβ+), double positive 3 (DP3; CD4+CD8+TCRβ+CD69+CD127+), CD4loCD8+ (CD4loCD8+; CD4loCD8+TCRβ+), CD8+ mature (CD8+ Mat; CD4-CD8+TCRβhiCD69-). See complete FACS gating strategy in Figure S5B. **(B)** Experimental overview of neonatal thymic slices cultured with a calcineurin inhibitor, Cyclosporin A (CsA). Postnatal day 1 (P1) thymic slices were harvested from mice and cultured at time point 0 in media alone containing no CsA or with various concentrations of CsA. Thymic slices were collected at indicated time points and analyzed via flow cytometry to quantify cell populations. Illustrations were created using Biorender.com. **(C)** Frequency (% of live cells) of CD4+CD8lo (CD4+CD8loTCRβ+) cells in slices from wild-type (WT; circles), MHCII-specific (β2M-/-; squares) or MHCI-specific (MHCII-/-; triangles) mice following culture in basal medium alone (No CsA; filled symbols) or with CsA (open symbols) for 96 hours. Data is compiled from 3 independent experiments with WT (B6) slices, 2 independent experiments with MHCII-specific slices (β2M-/-), and 5 independent experiments with MHCI-specific slices (MHCII-/-). Each symbol on the plots represents a thymic slice. Data was analyzed using an ordinary one-way ANOVA. **(D-G)** Frequency of **(D)** CD4+ SM, **(E)** CD4+ Mat, **(F)** CD4loCD8, and **(G)** CD8+ Mat cells from WT slices without CsA (filled symbols), as a control, or 200ng/mL CsA (open symbols). Data is compiled from 3 independent experiments with WT slices. Data was analyzed using an unpaired t test. **(H-I)** Frequency of **(H)** CD4+CD8lo or **(I)** CD4+ SM cells after 0, 24, 48, 72 and 96-hours of culture in basal medium alone (No CsA; filled symbols) or with 200ng/mL CsA (open symbols). Data is compiled from 8 independent experiments with WT slices. Data are displayed as mean ± standard error of the mean (SEM). For slices cultured with no CsA for 0 hours n=6, 24 hours n=6, 48 hours n=6, 72 hours n=22, 96 hours n=9. For slices cultured with 200ng/mL CsA for 24 hours n=3, 48 hours n=3, 72 hours n=15, 96 hours n=6. Data was analyzed using an ordinary two-way ANOVA with multiple comparisons. **(J-O)** GATA3 staining determined by using the FMO control. Data is from 1 experiment with WT slices. Frequency **(J)** and geometric mean fluorescent intensity (gMFI) **(K)** of GATA3+ Sig DP cells after 72 hours of culture in basal medium alone (No CsA; filled symbols) or with 200ng/mL CsA (open symbols). **(L)** Histogram displaying GATA3 expression of Sig DP cells cultured in basal medium alone (No CsA; solid, plum line) or with 200ng/mL CsA (dashed, plum line) and fluorescent minus one (FMO) control (solid, black line). **(M)** Frequency and **(N)** geometric mean fluorescent intensity (gMFI) of GATA3+ CD4+CD8lo cells after 72 hours of culture in basal medium alone (No CsA; filled symbols) or with 200ng/mL CsA (open symbols). **(O)** Histogram displaying Gata3 expression of CD4+CD8lo cells cultured in basal medium alone (No CsA; solid, pink line) or with 200ng/mL CsA (dashed, pink line) and fluorescent minus one (FMO) control (solid, black line). Data was analyzed using an unpaired t test. NS is not significant, *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

To gain insight into the timing of calcineurin-NFAT signaling in CD4 lineage commitment, we added CsA to neonatal slice cultures and tracked development every 24 hours over the course of 96 hours. We observed that the reduction in CD4+CD8lo cells when calcineurin-NFAT is inhibited was detectable after 48 hours in culture, and very apparent after 72 and 96 hours (Figure 5H). A reduction in CD4+ SM cells was observed slightly later after 72 and 96 hours (Figure 5I). These time course experiments confirm that inhibition of calcineurin-NFAT is affecting CD4+ T cells at early time points in lineage commitment.

GATA3 is a downstream target of NFAT and is required for CD4 lineage commitment (Pai et al., 2003; Aliahmad and Kaye, 2008; Lee et al., 2018). In addition, GATA3 is known to directly induce the expression of *Zbtb7b* (Wang et al., 2008). Therefore, we surmised that inhibition of calcineurin-NFAT may disrupt GATA3 expression, and thus block CD4 lineage choice by preventing sufficient expression of THPOK. To test this hypothesis, we measured GATA3 expression in neonatal slices cultured with CsA. While we did not observe a reduction in signaled double positive (Sig DP; CD4+CD8+CD69+) cells (Figure S5N), we did observe a reduction in GATA3+ Sig DP cells in the presence of CsA (Figure 5J), and a trend toward a reduction in

GATA3 expression (Figure 5K,L). In addition to a reduction of CD4+CD8lo cells, we also observed a reduction in the frequency of GATA3+ CD4+CD8lo (Figure 5M) cells. Of the few GATA3+ CD4+CD8lo cells present, we observed lower GATA3 expression in cells cultured with CsA compared to those cultured without (Figure 5N-O). These data support a requirement of calcineurin-NFAT early in CD4 lineage commitment, prior to GATA3 expression.

Discussion

In this study, we applied single-cell multi-omic analysis to investigate the development of thymocytes into CD4 and CD8 T cells. By measuring the transcriptome and surface proteins in developing thymocytes, we could define the continuous changes that occur over the course of this process, and relate these continuous changes to defined intermediate stages that could be isolated by fluorescence-based flow cytometry for further study. In particular, we clarified the progression of MHCI-specific thymocytes from early CD4+CD8+ DP stages to the CD4+CD8low stage, followed by a later CD4+CD8+ stage (DP3) before maturation into the CD8 SP stage. We also demonstrated that in silico gating of CITE-seq protein data can inspire gating strategies for fluorescence-based flow cytometry. While differences between sequencing-based and fluorescence-based protein measurements such as noise (e.g., spectral overlap) and sensitivity (e.g., barcode amplification) might limit the direct translation of gate position, large CITE-seq panels could provide a useful platform for screening potential combinations of fluorescence-based markers. By performing multi-omic measurements on thymocytes from both WT and lineage-restricted mice, we were able to identify the relative order and timing of key lineage-specifying differences. Our analysis bolstered support for the importance of early differences in TCR signaling in CD4/CD8 lineage commitment and generated the hypothesis that TCR signaling specifically through the calcineurin-NFAT axis provides the basis for early commitment into the CD4 rather than the CD8 lineage. Pharmacological inhibition of calcineurin-NFAT signaling with Cyclosporin A in *ex vivo* thymic slice culture supported this hypothesis, establishing the role of the calcineurin-NFAT axis of TCR signaling in CD4/CD8 lineage commitment.

By synthesizing our findings on RNA and protein expression event timing from CITE-seq data with the chronological timing of population phenotypes in *ex vivo* thymic slice cultures (Figure S4), we could construct a temporal model for CD4 versus CD8 T cell lineage commitment. Our data suggest that all positively-selected DP thymocytes begin the process of lineage commitment by "auditioning" for the CD4 lineage. During this early CD4 auditioning phase of positive selection, most MHCII-specific thymocytes receive moderate, persistent TCR signals, allowing them to lock in the CD4 fate by fully upregulating THPOK, activating the THPOK positive autoregulation loop (Muroi et al., 2008), leading to repression of CD8 and RUNX3. In contrast, most MHCI-specific thymocytes receive weaker, more transient signals during this phase (due to the combined effect of weaker LCK recruitment by the CD8 co-receptor, drop in CD8 surface expression and increase in the negative regulator CD5 (Chan et al., 1999). During the later CD8 lineage specification window, the CD8 SP enhancer and RUNX3 are activated, likely due to the maturation-associated drop in E protein activity (Jones-Mason et al., 2012). Continued upregulation of RUNX3 represses both THPOK and CD4, while further enhancing CD8 expression. In addition, an increase in TCR sensitivity during the CD8 lineage specification phase (due to the drop in negative regulator CD5, rise in ZAP70 (ref), and increase in ion channel components (e.g., KCNA2 and TMIE (Lutes et al., 2021)) leads to a second wave of TCR signaling. TCR signaling at this late stage of CD8 specification may provide survival signals and

well as further upregulating RUNX3, and would serve to ensure the elimination of any MHCII-specific thymocytes that failed the CD4 audition phase.

While in this study we focused our analysis on the CD4/CD8 lineage commitment, we anticipate that our approach using single-cell RNA and protein data could be applied to the analysis of other developmental systems such as the selection of Tregs within the thymus and the commitment of naive T cells to specialized functions within the periphery. The simultaneous measurement of RNA and protein not only allowed us to track the differences in relative timing of RNA and protein expression events, but also enabled the direct connection between multi-omic cell profiles and tangible populations that could be gated by flow cytometry for further analysis, such as intracellular transcription factor staining. The CITE-seq method is currently limited to the measurement of surface proteins (Stoeckius et al., 2017), but future technological developments to facilitate the simultaneous measurement of RNA, surface proteins, and large panels of intracellular proteins could greatly enhance the ability to generate hypotheses about molecular pathway activity, gene regulatory networks, and transcription and translation dynamics. Furthermore, because thymocytes actively traverse the thymic cortex and medulla over the course of their development, imaging could provide a valuable dimension to our current understanding of the thymocyte developmental timeline (Germain et al., 2012). Future work that integrates spatial locations with the transcriptomic and surface protein profiles of this study could inform how a cell's environment and physical motility might influence and reflect key aspects of the thymocyte developmental trajectory.

Acknowledgments

We thank BioLegend Inc. and their proteogenomics team, especially Kristopher Nazor, Bertrand Yeung, Andre Fernandes, Qing Gao, Hong Zhang, and Tse Shun Huang, for providing reagents and expertise and for help with sample preparation, library generation, and sequencing for a portion of the CITE-seq libraries used in this study. We thank the Flow Cytometry Core Facilities at UC Berkeley, including Hector Nolla and Alma Valeros for their help operating cell sorters. We thank the UC Berkeley Functional Genomics Lab, especially Justin Choi. We thank Silvia Ariotti for insightful early discussions, and Adam Gayoso for helpful discussions on the application of totalVI. We thank members of the Streets, Yosef, and Robey laboratories for helpful feedback. Research reported in this manuscript was supported by the NIGMS of the National Institutes of Health under award number R35GM124916 (A.S), the Chan Zuckerberg Foundation Network under grant number 2019-02452 (N.Y.) and the National Institutes of Mental Health under grant number U19MH114821 (N.Y.). Z.S. was supported by the National Science Foundation Graduate Research Fellowship and the Siebel Scholars award. N.Y. was supported by the Koret-Berkeley-Tel Aviv Initiative in Computational Biology. A.S. and N.Y. are Chan Zuckerberg Biohub investigators.

Author Contributions

Z.S., E.A.R., A.S., and N.Y. conceived the work. Z.S. and L.K.L. performed CITE-seq experiments. L.L.M. performed thymic slice and flow cytometry experiments. Z.S. designed and implemented analysis methods. L.L.M. and E.A.R. analyzed flow cytometry data with input from Z.S. Z.S. wrote the original draft. All authors reviewed and edited the manuscript. E.A.R., A.S., and N.Y. supervised the work.

Declaration of Interests
The authors declare no competing interests.

Methods

CITE-seq on mouse thymocytes
Mice: Wild type B6 (C57BL/6, Stock No.: 000664), B2M-/- (B6.129P2-B2m^{tm1Unc}/DcrJ, Stock No.: 002087), OT-I (C57BL/6-Tg(TcraTcrb)1100Mjb/J, Stock No.: 003831), and OT-II (B6.Cg-Tg(TcraTcrb)425Cbn/J, Stock No.: 004194) were obtained from The Jackson Laboratory. MHCII-/- (I-Aβ-/-) mice have been previously described (Grusby et al., 1991). RAG1-/-AND TCRtg mice and RAG1-/-F5 TCRtg were generated by crossing AND TCRtg (B10.Cg-Tg(TcrAND)53Hed/J, Jax Stock No.: 002761; (Kaye et al., 1989)) and F5 TCRtg (C57BL/6-Tg(CD2-TcraF5,CD2-TcrbF5)1Kio; (Mamalaki et al., 1992)) mice with RAG1-/- mice (Rag1-/-B6.129S7-Rag1tm1Mom) as previously described by (Au-Yeung et al., 2014)). All mice used in CITE-seq experiments were females between four and eight weeks of age. Samples are further described in Table S1. Mice were group housed with enrichment and segregated by sex in standard cages on ventilated racks at an ambient temperature of 26 °C and 40% humidity. Mice were kept in a dark/light cycle of 12h on and 12h off and given access to food and water ad libitum. All animal care and procedures were carried out in accordance with guidelines approved by the Institutional Animal Care and Use Committees at the University of California, Berkeley and at BioLegend, Inc.

Cell preparation: Mice were sacrificed, and thymi were harvested, placed in RPMI + 10% FBS medium on ice, mechanically dissociated with a syringe plunger, and passed through a 70 μm strainer to generate a single-cell suspension.

Antibody panel preparation: We prepared a panel containing 111 antibodies (TotalSeq-A mouse antibody panel 1, BioLegend, 900003217), which are enumerated in Table S2. Immediately prior to cell staining, we centrifuged the antibody panel for 10 minutes at 14,000g to remove antibody aggregates. We then performed a buffer exchange on the supernatant using a 50 kDa Amicon spin column (Millipore, UFC505096) following the manufacturer's protocol to transfer antibodies into RPMI + 10% FBS.

Cell sorting: To enrich for positively-selecting thymocytes in MHC-deficient and some wild-type samples (Table S1), live, single, TCRb+CD5+ thymocytes were sorted by FACS. We took advantage of the fact that cells were already stained with TotalSeq (oligonucleotide-conjugated) antibodies and therefore designed oligonucleotide-fluorophore conjugates complementary to the TotalSeq barcodes (5'-CACTGAGCTGTGGAA-AlexaFluor488-3' for CD5; 5'-TCCCATAGGATGGAA-AlexaFluor647-3' for TCRb). Prior to cell staining, the TotalSeq antibody panel was mixed with oligonucleotide-fluorophore conjugates in a 1:1.5 molar ratio. This mixture was incubated for 15 minutes at room temperature to allow for oligonucleotide hybridization, and then transferred to ice. Cells were then stained with the antibody/oligonucleotide-fluorophore mixture according to the TotalSeq protocol. Cells were stained, washed, and resuspended in RPMI + 10% FBS to maintain viability. Cells were sorted using a BD FACSAria Fusion (BD Biosciences).

CITE-seq protocol and library preparation: The CITE-seq experiment was performed following the TotalSeq protocol. Cells were stained, washed, and resuspended in RPMI + 10% FBS to maintain viability. We followed the 10X Genomics Chromium Single Cell 3′ v3 protocol to prepare RNA and antibody-derived-tag (ADT) libraries (Zheng et al., 2017).

Sequencing and data processing: RNA and ADT libraries were sequenced with either an Illumina NovaSeq S1 or an Illumina NovaSeq S4. Reads were processed with Cell Ranger v.3.1.0 with feature barcoding, where RNA reads were mapped to the mouse mm10–2.1.0 reference (10X Genomics, STAR aligner (Dobin et al., 2013)) and antibody reads were mapped to known barcodes (Table S1). No read depth normalization was applied when aggregating samples.

CITE-seq data preprocessing

Prior to analysis with totalVI, we performed preliminary quality control and feature selection on the CITE-seq data. Cells with a high percentage of UMIs from mitochondrial genes (> 15% of a cell's total UMI count) were removed. We also removed cells expressing < 200 genes, and retained cells with protein library size between 1,000 and 10,000 UMI counts. We removed cells in which fewer than 70 proteins were detected of the 111 measured in the panel. An initial gene filter removed genes expressed in fewer than four cells. The top 5,000 highly variable genes (HVGs) were selected by the Seurat v3 method (Stuart et al., 2019) as implemented by scVI (Lopez et al., 2018). In addition to HVGs, we also selected genes encoding proteins in the measured antibody panel and a manually selected set of genes of interest. After all filtering, the CITE-seq dataset contained a total of 72,042 cells, 5,125 genes, and 111 proteins.

CITE-seq data analysis with totalVI

totalVI modeling of all CITE-seq data: We ran totalVI on CITE-seq data after filtering (described above), using a 20-dimensional latent space. Each 10X lane was treated as a batch. When generating denoised gene and protein values, we applied the *transform_batch* parameter (Gayoso et al., 2021) to view all denoised values in the context of wild-type samples.

Cell annotation: We stratified cells of the thymus into cell types and states based on the totalVI latent space, taking advantage of both RNA and protein information. We first clustered cells in the totalVI latent space with the Scanpy (Wolf et al., 2018) implementation of the Leiden algorithm (Traag et al., 2019) at resolution 0.6, resulting in 18 clusters. We repeated this approach to subcluster cells. We used Vision (DeTomaso et al., 2019) with default parameters for data exploration. Subclusters were manually annotated based on curated lists of cell type markers (Gayoso et al., 2021; Hogquist et al., 2015), resulting in 20 annotated clusters (excluding one cluster annotated as doublets). We visualized the totalVI latent space in two dimensions using the Scanpy (Wolf et al., 2018) implementation of the UMAP algorithm (Becht et al., 2019).

Differential expression testing of annotated cell types: We conducted a one-vs-all differential expression test between all annotated cell types, excluded clusters annotated as doublets or dying cells. We identified cell type markers by filtering for significance (log(Bayes factor) > 2.0 for genes, log(Bayes factor) > 1.0 for proteins), effect size (median log fold change (LFC) > 0.2 for both genes and proteins), and the proportion of expressing cells (detected expression in > 10% of the relevant population for genes), and sorting by the median LFC. For marker visualization, we

selected the top four (if existing) differentially expressed genes and proteins per cell type, arranged by the cell type in which the LFC was highest.

totalVI modeling of positively-selecting thymocytes: To further analyze thymocyte populations with a focus on positively-selected cells, we selected the following annotated clusters: Signaled DP, Immature CD4, Immature CD8, Mature CD4, Mature CD8, Interferon signature cells, Negative selection (wave 2), and Treg. With an interest in the variation within thymocyte populations (rather than all cells in the thymus), we selected the top 5,000 HVGs in this subset, as well as genes encoding proteins in the measured antibody panel and a manually selected set of genes of interest. This resulted in a CITE-seq dataset containing 35,943 cells, 5,108 genes, and 111 proteins. We ran totalVI on this subset dataset and generated denoised values as described above. We performed Leiden clustering and visualized the totalVI latent space in two dimensions using UMAP as described above.

Cell filtering of positively-selecting thymocytes on the CD4/CD8 developmental trajectory: After visualizing the totalVI latent space of the thymocyte subset, we applied additional filters to restrict to cells on the CD4/CD8 developmental trajectory. We used two resolutions of Leiden clustering (0.6 and 1.4) and subclustering as described above to identify and remove clusters of negatively selected cells, Tregs, gamma-delta-like cells, mature cycling cells, and outlier clusters of doublets, interferon-responding cells, and CD8-transgenic-specific cells. After filtering, this dataset contained 29,408 cells that were used for downstream analysis. Differential expression testing of positively-selecting thymocytes using pseudotime information is described below.

Pseudotime inference

Pseudotime inference with Slingshot: Slingshot (Street et al., 2018) was selected for pseudotime inference based on its superior performance in a comprehensive benchmarking study (Saelens et al., 2019). Slingshot pseudotime was derived from the UMAP projection of the totalVI latent space. The starting point was assigned to DP cells, and two endpoints were assigned to mature CD4 and CD8 T cells. Slingshot pseudotime derived from the full 20-dimensional totalVI latent space was highly correlated with that from the 2-dimensional space (Figure S2A), supporting our use of the 2D-derived pseudotime values for ease of visualization and analysis.

Lineage assignment: Initial lineage assignment of cells was made on the basis of their genotype (CD4 lineage for MHCII-/-, AND, and OT-II mice, CD8 lineage for MHCI-/-, F5, and OT-I mice, unassigned for B6 mice). However, small numbers of cells in MHC-deficient and TCR transgenic mice develop along the alternative lineage (particularly in TCR transgenics that are *Rag+*, which might express an endogenous TCR in addition to the transgenic TCR). We therefore added an additional filter of Slingshot lineage assignment weight > 0.5. Cells with a Slingshot lineage assignment weight of < 0.5 along the expected lineage based on genotype were excluded from the remaining pseudotime-based analysis.

In silico flow cytometry

To perform in silico flow cytometry, totalVI denoised protein counts were log-transformed and visualized in biaxial-style scatter plots. Gates in biaxial plots were determined based on contours of cell density. An approximate alignment of gated populations to pseudotime was generated by

identifying thresholds classifying adjacent populations in pseudotime by maximizing the Youden criteria.

Adult thymocyte population analysis with fluorescence-based flow cytometry

Mice: All experiments were approved by the University of California, Berkeley Animal Use and Care Committee. All mice were bred and maintained under pathogen-free conditions in an American Association of Laboratory Animal Care-approved facility at the University of California, Berkeley. Wild type B6 (C57BL/6, Stock No.: 000664) and B2M-/- (B6.129P2-B2m^{tm1Unc}/DcrJ, Stock No.: 002087) were obtained from The Jackson Laboratory. MHCII-/- (I-Aβ-/-) mice have been previously described (Grusby et al., 1991). For thymocyte population analysis in adult mice, six to eight week- old animals were used. Thymi were analyzed from four mice per genotype (2 male and 2 female).

Flow cytometry: Thymi were mechanically dissociated into a single-cell suspension, depleted of red blood cells using ACK Lysis Buffer (0.15M NH4CL, 1mM kHC3, 0.1mM Na2EDTA). Cells were filtered, washed, and counted before being stained with a live/dead stain; Zombie NIR Fixable Viability Kit (Biolegend). Samples were blocked with anti-CD16/32 (2.4G2) and surface antibodies against CD4, CD8, TCRβ, CD5, CD69, and CD127 (IL-7R). Intracellular staining for GATA3, RUNX3, and THPOK was performed using the eBioscience FOXP3/ Transcription Factor Staining Buffer Set (Thermo Fisher). All antibodies were purchased from BD Biosciences, Biolegend, or eBiosciences. Single-stain samples and fluorescence minus one (FMO) controls were used to establish PMT voltages, gating and compensation parameters. Cells were processed using a BD LSRFortessa or BD LSRFortessa X20 flow cytometer and analyzed using FlowJo software (Tree Star).

Differential expression analysis of positively-selecting thymocytes with totalVI

Testing for temporal features: Temporal features (i.e., features that are differentially expressed over time) were determined by a totalVI one-vs-all DE test within each lineage between binned units of pseudotime. DE criteria (as above) included filters for significance (log(Bayes factor) > 2.0 for genes, log(Bayes factor) > 1.0 for proteins), effect size (median log fold change > 0.2 for both genes and proteins), and the proportion of expressing cells (detected expression in > 5% of the relevant population for genes). Top temporal genes were selected as the unique set among the top three differentially expressed genes per time in each lineage.

Testing for differences between lineages: Differences between lineages were determined by a totalVI within-cluster DE test, where clusters were binned units in pseudotime and the condition was lineage assignment (i.e., cells within a given unit of pseudotime were compared between lineages). Criteria for DE were the same as above.

Clustering of differentially expressed genes: To cluster differentially expressed genes into patterns, totalVI denoised gene expression values were standard scaled, reduced dimensions across cells using PCA, and clustered genes using the Leiden algorithm (Traag et al., 2019) as implemented by Scanpy (Wolf et al., 2018). For temporal features, clustering was performed across all cells. For features differentially expressed between lineages, the genes upregulated within a lineage were clustered according to expression within the lineage in which they were upregulated.

Transcription factor enrichment analysis

ChEA3 analysis: To perform transcription factor enrichment analysis with ChEA3 (Keenan et al., 2019), we first selected target gene sets as genes differentially upregulated in one lineage relative to the other in each unit of pseudotime, filtered for significance (log(Bayes factor) > 2.0), effect size (median log fold change > 0.2), and detected expression in > 5% of the population of interest. For each target gene set, transcription factors (TFs) were scored for enrichment by the integrated mean ranking across all ChEA3 gene set libraries (MeanRank) based on the top performance of this ranking method (Keenan et al., 2019).

Ranking of candidate TFs: To generate an overall ranking of TFs for their likely involvement in CD4/CD8 lineage commitment, we focused on enrichment in the three units of pseudotime prior to master regulator differential expression in each lineage (i.e., in the CD4 lineage, the relevant pseudotime units are 4, 5, and 6 prior to the differential expression of *Zbtb7b* differential expression at pseudotime 7; in the CD8 lineage, the relevant pseudotime units are 5, 6, and 7 prior to the differential expression of *Runx3* at pseudotime 8). We excluded the pseudotime unit containing master regulator differential expression from the ranking, as genes differentially expressed at this time could be the result of the master regulator itself enforcing lineage-specific changes rather than the factors driving initial commitment to a lineage. The pseudotime unit containing master regulator differential expression is included in Figure 4F-G for visualization, but did not contribute to the ranked order of TFs. We also excluded earlier units of pseudotime since these times included very few (< 15) significantly different genes between the lineages. Finally, we required that the TF itself had detected expression in at least 5% of the relevant population. The overall ranking of candidate driver TFs was then generated by taking the mean of ranks across the relevant pseudotime units.

TCR signaling pathway involvement: TFs were annotated by whether they had a known association with TCR signaling. A list of molecules involved in TCR signaling were curated from the NetPath database of molecules involved in the TCR signaling pathway and the NetPath database of genes transcriptionally upregulated by the TCR signaling pathway (Kandasamy et al., 2010). Additional genes related to TCR signaling were curated from literature sources (Shao et al., 1997; Wong et al., 2014; Lopez-Rodriguez et al., 2015; Hedrick et al., 2013; Wang et al., 2010). TFs were also annotated by whether they were known to target either *Gata3, Zbtb7b,* or *Runx3* according to ChEA3 databases (i.e., *Gata3, Zbtb7b,* or *Runx3* appeared in the Overlapping Gene list for the TF of interest in any ChEA3 query).

Neonatal thymic slice experiments

Mice: All experiments were approved by the University of California, Berkeley Animal Use and Care Committee. All mice were bred and maintained under pathogen-free conditions in an American Association of Laboratory Animal Care-approved facility at the University of California, Berkeley. Wild type B6 (C57BL/6, Stock No.: 000664) and β2M-/- (B6.129P2-B2m^{tm1Unc}/DcrJ, Stock No.: 002087) were obtained from The Jackson Laboratory. MHCII-/- (I-Aβ-/-) mice have been previously described (Grusby et al., 1991). For neonatal thymic slice experiments, postnatal day 1 (P1) mice were used.

Thymic slices: Thymic slices were prepared as previously described (Dzhagalov et al., 2012; Ross et al., 2016), with minor modifications to adjust for the smaller size of neonatal thymi compared

to those of adults. Thymic lobes were dissected, removed of connective tissue, embedded in 4% low melting point agarose (GTG-NuSieve Agarose, Lonza) and sectioned into 500 μM slices using a vibratome (VT1000S, Leica). Slices were overlaid onto 0.4 μM transwell inserts (Corning, Cat. No.: 353090) and placed in a 6-well tissue culture plate with 1 mL of complete RPMI medium (RPMI-1640 (Corning), 10% FBS (Thermo), 100U/mL penicillin/streptomycin (Gibco), 1X L-glutamine (Gibco), 55μM 2-mercaptoethanol (Gibco)). Slices were cultured for indicated periods of time at 37 °C, 5% CO_2, before being prepared and analyzed by flow cytometry. For neonatal slice cultures containing Cyclosporin A (CsA; Millipore-Sigma, Cat. No.:239835), CsA was serially diluted to indicated concentrations (50-800ng/mL) and added directly to the culture medium.

Flow cytometry: Thymic slices were mechanically dissociated into a single-cell suspension, depleted of red blood cells using ACK Lysis Buffer. Cells were filtered, washed and counted before being stained with a live dead/stain; Propidium Iodine (Biolegend), Ghost Violet 510 (Tonbo), Zombie NIR, or Zombie UV Fixable Viability Kit (Biolegend). Samples were blocked with anti-CD16/32 (2.4G2) and stained with surface antibodies against CD4, CD8, TCRβ, and CD69. Intracellular staining for GATA3, RUNX3, and THPOK was performed using the eBioscience FoxP3/ Transcription Factor Staining Buffer Set (Thermo Fisher). All antibodies were purchased from BD Biosciences, Biolegend, or eBiosciences. Single-stain samples and fluorescence minus one (FMO) controls were used to establish PMT voltages, gating and compensation parameters. Cells were processed using a BD LSRFortessa or BD LSRFortessa X20 flow cytometer and analyzed using FlowJo software (Tree Star).

Statistical analysis: Data were analyzed using Prism software (GraphPad). Comparisons were performed using an unpaired T test, one- or two-way analysis of variance, where indicated in the figure legends. For all statistical models and tests described above, the significance is displayed as follows; ns is not significant, *$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$.

Supplemental Tables

Table S1: CITE-seq sample information.
Table S2: Antibodies used in this study.
Table S3: DE test results for totalVI one-versus-all DE test between annotated thymus populations.
Table S4: DE test results for totalVI DE test across pseudotime within the CD4 lineage.
Table S5: DE test results for totalVI DE test across pseudotime within the CD8 lineage.
Table S6: Cluster assignments for genes with temporal patterns from the totalVI DE tests across pseudotime.
Table S7: DE test results for totalVI DE test within pseudotime and between CD4 and CD8 lineages.
Table S8: Cluster assignments for genes upregulated in the CD4 lineage from the totalVI DE test within pseudotime and between CD4 and CD8 lineages.
Table S9: Cluster assignments for genes upregulated in the CD8 lineage from the totalVI DE test within pseudotime and between CD4 and CD8 lineages.

Data and Code Availability

CITE-seq data are being uploaded to GEO. An accession number will be provided once available. Code will be made available upon request.

References

Aliahmad, P., & Kaye, J. (2008). Development of all CD4 T lineages requires nuclear factor TOX. *Journal of Experimental Medicine*, *205*(1), 245–256. https://doi.org/10.1084/jem.20071944

Au-Yeung, B. B., Melichar, H. J., Ross, J. O., Cheng, D. A., Zikherman, J., Shokat, K. M., … Weiss, A. (2014). Quantitative and temporal requirements revealed for Zap70 catalytic activity during T cell development. *Nature Immunology*, *15*(7), 687–694. https://doi.org/10.1038/ni.2918

Azzam, H. S., Grinberg, A., Lui, K., Shen, H., Shores, E. W., & Love, P. E. (1998). CD5 expression is developmentally regulated by T cell receptor (TCR) signals and TCR avidity. *Journal of Experimental Medicine*, *188*(12), 2301–2311. https://doi.org/10.1084/jem.188.12.2301

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., … Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, *37*(1), 38–44. https://doi.org/10.1038/nbt.4314

Bosselut, R. (2004). CD4/CD8-lineage differentiation in the thymus: From nuclear effectors to membrane signals. *Nature Reviews Immunology*. Nature Publishing Group. https://doi.org/10.1038/nri1392

Chakraborty, A. K., & Weiss, A. (2014). Insights into the initiation of TCR signaling. *Nature Immunology*. Nature Publishing Group. https://doi.org/10.1038/ni.2940

Chan, S., Waltzinger, C., Tarakhovsky, A., Benoist, C., & Mathis, D. (1999). An influence of CD5 on the selection of CD4-lineage T cells. *European Journal of Immunology*, *29*(9), 2916–2922. https://doi.org/10.1002/(SICI)1521-4141(199909)29:09<2916::AID-IMMU2916>3.0.CO;2-I

Chèneby, J., Ménétrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., … Ballester, B. (2020). ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Research*, *48*(D1), D180–D188. https://doi.org/10.1093/nar/gkz945

Chopp, L. B., Gopalan, V., Ciucci, T., Ruchinskas, A., Rae, Z., Lagarde, M., … Bosselut, R. (2020). An Integrated Epigenomic and Transcriptomic Map of Mouse and Human ab T Cell Development. *Immunity*, *53*(6). https://doi.org/10.1016/j.immuni.2020.10.024

Daley, S. R., Hu, D. Y., & Goodnow, C. C. (2013). Helios marks strongly autoreactive CD4+ T cells in two major waves of thymic deletion distinguished by induction of PD-1 or NF-κB. *Journal of Experimental Medicine*, *210*(2), 269–285. https://doi.org/10.1084/jem.20121458

Daniels, M. A., Teixeiro, E., Gill, J., Hausmann, B., Roubaty, D., Holmberg, K., … Palmer, E. (2006). Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling. *Nature*, *444*(7120), 724–729. https://doi.org/10.1038/nature05269

DeTomaso, D., Jones, M. G., Subramaniam, M., Ashuach, T., Ye, C. J., & Yosef, N. (2019). Functional interpretation of single cell similarity maps. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12235-0

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Dzhagalov, I. L., Melichar, H. J., Ross, J. O., Herzmark, P., & Robey, E. A. (2012). Two-Photon Imaging of the Immune System. In *Current Protocols in Cytometry* (Vol. 60, pp. 12.26.1-12.26.20). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471142956.cy1226s60

Gallo, E. M., Winslow, M. M., Canté-Barrett, K., Radermacher, A. N., Ho, L., McGinnis, L., … Crabtree, G. R. (2007). Calcineurin sets the bandwidth for discrimination of signals during thymocyte development. *Nature*, *450*(7170), 731–735. https://doi.org/10.1038/nature06305

Gao, E. K., Lo, D., Cheney, R., Kanagawa, O., & Sprent, J. (1988). Abnormal differentiation of thymocytes in mice treated with cyclosporin A. *Nature*, *336*(6195), 176–179. https://doi.org/10.1038/336176a0

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, *18*(3). https://doi.org/10.1038/s41592-020-01050-x

Germain, R. N. (2002). t-cell development and the CD4-CD8 lineage decision. *Nature Reviews Immunology*. European Association for Cardio-Thoracic Surgery. https://doi.org/10.1038/nri798

Germain, R. N., Robey, E. A., & Cahalan, M. D. (2012, June 29). A Decade of imaging cellular motility and interaction dynamics in the immune system. *Science*. American Association for the Advancement of Science. https://doi.org/10.1126/science.1221063

Grusby, M. J., Johnson, R. S., Papaioannou, V. E., & Glimcher, L. H. (1991). Depletion of CD4+ T cells in major histocompatibility complex class II-deficient mice. *Science*, *253*(5026), 1417–1420. https://doi.org/10.1126/science.1910207

Hedrick, S. M., Michelini, R. H., Doedens, A. L., Goldrath, A. W., & Stone, E. L. (2012, September). FOXO transcription factors throughout T cell biology. *Nature Reviews Immunology*. NIH Public Access. https://doi.org/10.1038/nri3278

Hettmann, T., & Leiden, J. M. (2000). NF-κB Is Required for the Positive Selection of CD8 + Thymocytes . *The Journal of Immunology*, *165*(9), 5004–5010. https://doi.org/10.4049/jimmunol.165.9.5004

Hogquist, K. A., & Jameson, S. C. (2014). The self-obsession of T cells: How TCR signaling thresholds affect fate "decisions" and effector function. *Nature Immunology*, *15*(9), 815–823. https://doi.org/10.1038/ni.2938

Hogquist, K., Xing, Y., Hsu, F.-C., & Shapiro, V. S. (2015). T Cell Adolescence: Maturation Events Beyond Positive Selection. *Journal of Immunology*, *195*(4), 1351–1357. https://doi.org/10.4049/jimmunol.1501050

Jenkins, M. K., Schwartz, R. H., & Pardoll, D. M. (1988). Effects of cyclosporine A on T cell development and clonal deletion. *Science*, *241*(4873), 1655–1658. https://doi.org/10.1126/science.3262237

Jimi, E., Strickland, I., Voll, R. E., Long, M., & Ghosh, S. (2008). Differential Role of the Transcription Factor NF-κB in Selection and Survival of CD4+ and CD8+ Thymocytes. *Immunity*, *29*(4), 523–537. https://doi.org/10.1016/j.immuni.2008.08.010

Jones-Mason, M. E., Zhao, X., Kappes, D., Lasorella, A., Iavarone, A., & Zhuang, Y. (2012). E Protein Transcription Factors Are Required for the Development of CD4 + Lineage T Cells. *Immunity*, *36*(3), 348–361. https://doi.org/10.1016/j.immuni.2012.02.010

Kandasamy, K., Sujatha Mohan, S., Raju, R., Keerthikumar, S., Sameer Kumar, G. S., Venugopal, A. K., … Pandey, A. (2010). NetPath: A public resource of curated signal transduction pathways. *Genome Biology*, *11*(1), R3. https://doi.org/10.1186/gb-2010-11-1-r3

Kaye, J., Hsu, M. L., Sauron, M. E., Jameson, S. C., Gascoigne, N. R. J., & Hedrick, S. M. (1989). Selective development of CD4+ T cells in transgenic mice expressing a class II MHC-restricted antigen receptor. *Nature*, *341*(6244), 746–749. https://doi.org/10.1038/341746a0

Keenan, A. B., Torre, D., Lachmann, A., Leong, A. K., Wojciechowicz, M. L., Utti, V., … Ma'ayan, A. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Research*, *47*(W1), W212–W224. https://doi.org/10.1093/nar/gkz446

Kernfeld, E. M., Genga, R. M. J., Neherin, K., Magaletta, M. E., Xu, P., & Maehr, R. (2018). A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity*, *48*(6), 1258-1270.e6. https://doi.org/10.1016/j.immuni.2018.04.015

Kurd, N., & Robey, E. A. (2016). T-cell selection in the thymus: A spatial and temporal perspective. *Immunological Reviews*, *271*(1), 114–126. https://doi.org/10.1111/imr.12398

Lavaert, M., Liang, K. L., Vandamme, N., Park, J.-E., Roels, J., Kowalczyk, M. S., … Taghon, T. (2020). Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes. *Immunity*, *52*(6). https://doi.org/10.1016/j.immuni.2020.03.019

Lee, J. U., Kim, L. K., & Choi, J. M. (2018, November 27). Revisiting the concept of targeting NFAT to control T cell immunity and autoimmune diseases. *Frontiers in Immunology*. Frontiers Media S.A. https://doi.org/10.3389/fimmu.2018.02747

Liu, J. (1993). FK506 and cyclosporin, molecular probes for studying intracellular signal transduction. *Immunology Today*, *14*(6), 290–295. https://doi.org/10.1016/0167-5699(93)90048-P

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

López-Rodríguez, C., Aramburu, J., & Berga-Bolaños, R. (2015). Transcription factors and target genes of pre-TCR signaling. *Cellular and Molecular Life Sciences*, *72*(12), 2305–2321. https://doi.org/10.1007/s00018-015-1864-8

Lucas, B., & Germain, R. N. (1996). Unexpectedly complex regulation of CD4/CD8 coreceptor expression supports a revised model for CD4+CD8+ thymocyte differentiation. *Immunity*, *5*(5), 461–477. https://doi.org/10.1016/S1074-7613(00)80502-6

Lutes, L. K., Steier, Z., McIntyre, L. L., Pandey, S., Kaminski, J., Hoover, A. R., … Robey, E. A. (2021). T cell self-reactivity during thymic development dictates the timing of positive selection. *ELife*, *10*. https://doi.org/10.7554/eLife.65435

Malissen, B., Grégoire, C., Malissen, M., & Roncagalli, R. (2014). Integrative biology of T cell activation. *Nature Immunology*. Nature Publishing Group. https://doi.org/10.1038/ni.2959

Mamalaki, C., Norton, T., Tanaka, Y., Townsend, A. R., Chandler, P., Simpson, E., & Kioussis, D. (1992). Thymic depletion and peripheral activation of class I major histocompatibility complex-restricted T cells by soluble peptide in T-cell receptor transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(23), 11342–11346. https://doi.org/10.1073/pnas.89.23.11342

McNeil, L. K., Starr, T. K., & Hogquist, K. A. (2005). A requirement for sustained ERK signaling during thymocyte positive selection in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13574–13579. https://doi.org/10.1073/pnas.0505110102

Melichar, H. J., Ross, J. O., Herzmark, P., Hogquist, K. A., & Robey, E. A. (2013). Distinct Temporal Patterns of T Cell Receptor Signaling During Positive Versus Negative Selection in Situ. *Science Signaling*, *6*(297), ra92–ra92. https://doi.org/10.1126/scisignal.2004400

Muroi, S., Naoe, Y., Miyamoto, C., Akiyama, K., Ikawa, T., Masuda, K., … Taniuchi, I. (2008). Cascading suppression of transcriptional silencers by ThPOK seals helper T cell fate. *Nature Immunology*, *9*(10), 1113–1121. https://doi.org/10.1038/ni.1650

Navarro, M. N., & Cantrell, D. A. (2014, August 19). Serine-threonine kinases in TCR signaling. *Nature Immunology*. Nature Publishing Group. https://doi.org/10.1038/ni.2941

Pai, S. Y., Truitt, M. L., Ting, C. N., Leiden, J. M., Glimcher, L. H., & Ho, I. C. (2003). Critical Roles for Transcription Factor GATA-3 in Thymocyte Development. *Immunity*, *19*(6), 863–875. https://doi.org/10.1016/S1074-7613(03)00328-5

Park, J. E., Botting, R. A., Conde, C. D., Popescu, D. M., Lavaert, M., Kunz, D. J., … Teichmann, S. A. (2020). A cell atlas of human thymic development defines T cell repertoire formation. *Science*, *367*(6480). https://doi.org/10.1126/science.aay3224

Ross, J. O., Melichar, H. J., Au-Yeung, B. B., Herzmark, P., Weiss, A., & Robey, E. A. (2014). Distinct phases in the positive selection of CD8+ T cells distinguished by intrathymic migration and T-cell receptor signaling patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(25). https://doi.org/10.1073/pnas.1408482111

Ross, J. O., Melichar, H. J., Halkias, J., & Robey, E. A. (2015). Studying T cell development in thymic slices. In *T-Cell Development: Methods and Protocols* (Vol. 1323, pp. 131–140). Springer New York. https://doi.org/10.1007/978-1-4939-2809-5_11

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Saini, M., Sinclair, C., Marshall, D., Tolaini, M., Sakaguchi, S., & Seddon, B. (2010). Regulation of Zap70 expression during thymocyte development enables temporal separation of CD4 and CD8 repertoire selection at different signaling thresholds. *Science Signaling*, *3*(114), ra23–ra23. https://doi.org/10.1126/scisignal.2000702

Shao, H., Kono, D. H., Chen, L. Y., Rubin, E. M., & Kaye, J. (1997). Induction of the early growth response (Egr) family of transcription factors during thymic selection. *Journal of Experimental Medicine*, *185*(4), 731–744. https://doi.org/10.1084/jem.185.4.731

Sharp, L. L., Schwarz, D. A., Bott, C. M., Marshall, C. J., & Hedrick, S. M. (1997). The influence of the MAPK pathway on T cell lineage commitment. *Immunity*, *7*(5), 609–618. https://doi.org/10.1016/S1074-7613(00)80382-9

Shi, Y., Sahai, B. M., & Green, D. R. (1989). Cyclosporin A inhibits activation-induced cell death in T-cell hybridomas and thymocytes. *Nature*, *339*(6226), 625–626. https://doi.org/10.1038/339625a0

Singer, A., Adoro, S., & Park, J. H. (2008, October). Lineage fate and intense debate: Myths, models and mechanisms of CD4- versus CD8-lineage choice. *Nature Reviews Immunology*. Nat Rev Immunol. https://doi.org/10.1038/nri2416

Steinke, F. C., Yu, S., Zhou, X., He, B., Yang, W., Zhou, B., … Xue, H. H. (2014). TCF-1 and LEF-1 act upstream of Th-POK to promote the CD4 + T cell fate and interact with Runx3 to silence Cd4 in CD8 + T cells. *Nature Immunology*, *15*(7), 646–656. https://doi.org/10.1038/ni.2897

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., … Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868. https://doi.org/10.1038/nmeth.4380

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., … Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*(1), 1–16. https://doi.org/10.1186/s12864-018-4772-0

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., … Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031

Taniuchi, I. (2016). Views on helper/cytotoxic lineage choice from a bottom-up approach. *Immunological Reviews*, *271*(1), 98–113. https://doi.org/10.1111/imr.12401

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-41695-z

Urdahl, K. B., Pardoll, D. M., & Jenkins, M. K. (1994). Cyclosporin A inhibits positive selection and delays negative selection in αβ TCR transgenic mice. *Journal of Immunology (Baltimore, Md. : 1950)*, *152*(6), 2853–289.

Vacchio, M. S., & Bosselut, R. (2016). Function Transcriptional Circuitry To Control Their Lineage Commitment + CD8 − + in the Thymus: How T Cells Recycle the CD4 What Happens in the Thymus Does Not Stay. *J Immunol References*, *196*, 4848–4856. https://doi.org/10.4049/jimmunol.1600415

Wang, C. R., Hashimoto, K., Kubo, S., Yokochi, T., Kubo, M., Suzuki, M., … Nakayama, T. (1995). T cell receptor-mediated signaling events in CD4+CD8+ thymocytes undergoing thymic selection: Requirement of calcineurin activation for thymic positive selection but not negative selection. *Journal of Experimental Medicine*, *181*(3), 927–941. https://doi.org/10.1084/jem.181.3.927

Wang, L., Wildt, K. F., Zhu, J., Zhang, X., Feigenbaum, L., Tessarollo, L., … Bosselut, R. (2008). Distinct functions for the transcription factors GATA-3 and ThPOK during intrathymic differentiation of CD4+ T cells. *Nature Immunology*, *9*(10), 1122–1130. https://doi.org/10.1038/ni.1647

Wang, L., Xiong, Y., & Bosselut, R. (2010, October 1). Tenuous paths in unexplored territory: From T cell receptor signaling to effector gene expression during thymocyte selection. *Seminars in Immunology*. Academic Press. https://doi.org/10.1016/j.smim.2010.04.013

Weist, B. M., Kurd, N., Boussier, J., Chan, S. W., & Robey, E. A. (2015). Thymic regulatory T cell niche size is dictated by limiting IL-2 from antigen-bearing dendritic cells and feedback competition. *Nature Immunology*, *16*(6), 635–641. https://doi.org/10.1038/ni.3171

Wilkinson, B., & Kaye, J. (2001). Requirement for sustained MAPK signaling in both CD4 and CD8 lineage commitment: A threshold model. *Cellular Immunology*, *211*(2), 86–95. https://doi.org/10.1006/cimm.2001.1827

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*. https://doi.org/10.1186/s13059-017-1382-0

Wong, W. F., Looi, C. Y., Kon, S., Movahed, E., Funaki, T., Chang, L. Y., … Kohu, K. (2014). T-cell receptor signaling induces proximal Runx1 transactivation via a calcineurin-NFAT pathway. *European Journal of Immunology*, *44*(3), 894–904. https://doi.org/10.1002/eji.201343496

Xing, Y., Wang, X., Jameson, S. C., & Hogquist, K. A. (2016). Late stages of T cell maturation in the thymus involve NF-κB and tonic type i interferon signaling. *Nature Immunology*, *17*(5), 565–573. https://doi.org/10.1038/ni.3419

Zamisch, M., Tian, L., Grenningloh, R., Xiong, Y., Wildt, K. F., Ehlers, M., … Bosselut, R. (2009). The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8

T cell differentiation in the thymus. *Journal of Experimental Medicine*, *206*(12), 2685–2699. https://doi.org/10.1084/jem.20092024

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*, 14049. https://doi.org/10.1038/ncomms14049

Zhou, W., Yui, M. A., Williams, B. A., Yun, J., Wold, B. J., Cai, L., & Rothenberg, E. V. (2019). Single-Cell Analysis Reveals Regulatory Gene Expression Dynamics Leading to Lineage Commitment in Early T Cell Development. *Cell Systems*, *9*(4), 321-337.e9. https://doi.org/10.1016/j.cels.2019.09.008

Supplemental figures

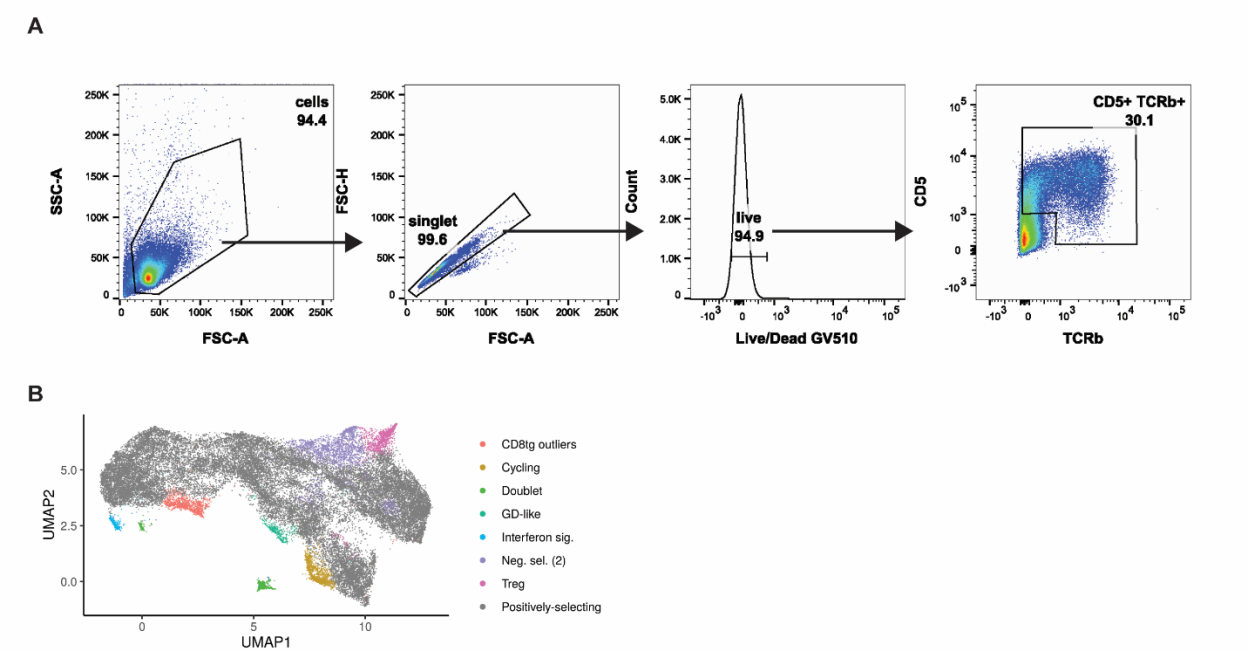


Figure S1: FACS sorting and cell filtering to enrich for positively-selecting thymocytes. (A) Representative FACS plots displaying gating strategy to sort thymocytes for CITE-seq. Cell populations were gated and sorted to include lymphocytes, exclude forward scatter doublets, include Ghost Dye Violet 510 Live/Dead stain negative (live cells), then on TCRβ+CD5+ to select for cells that were positively-selecting. **(B)** UMAP plot of totalVI latent space from positively-selected thymocytes before filtering indicating annotated populations that were retained (positively-selecting thymocytes)) or removed (all other populations) from downstream analysis.

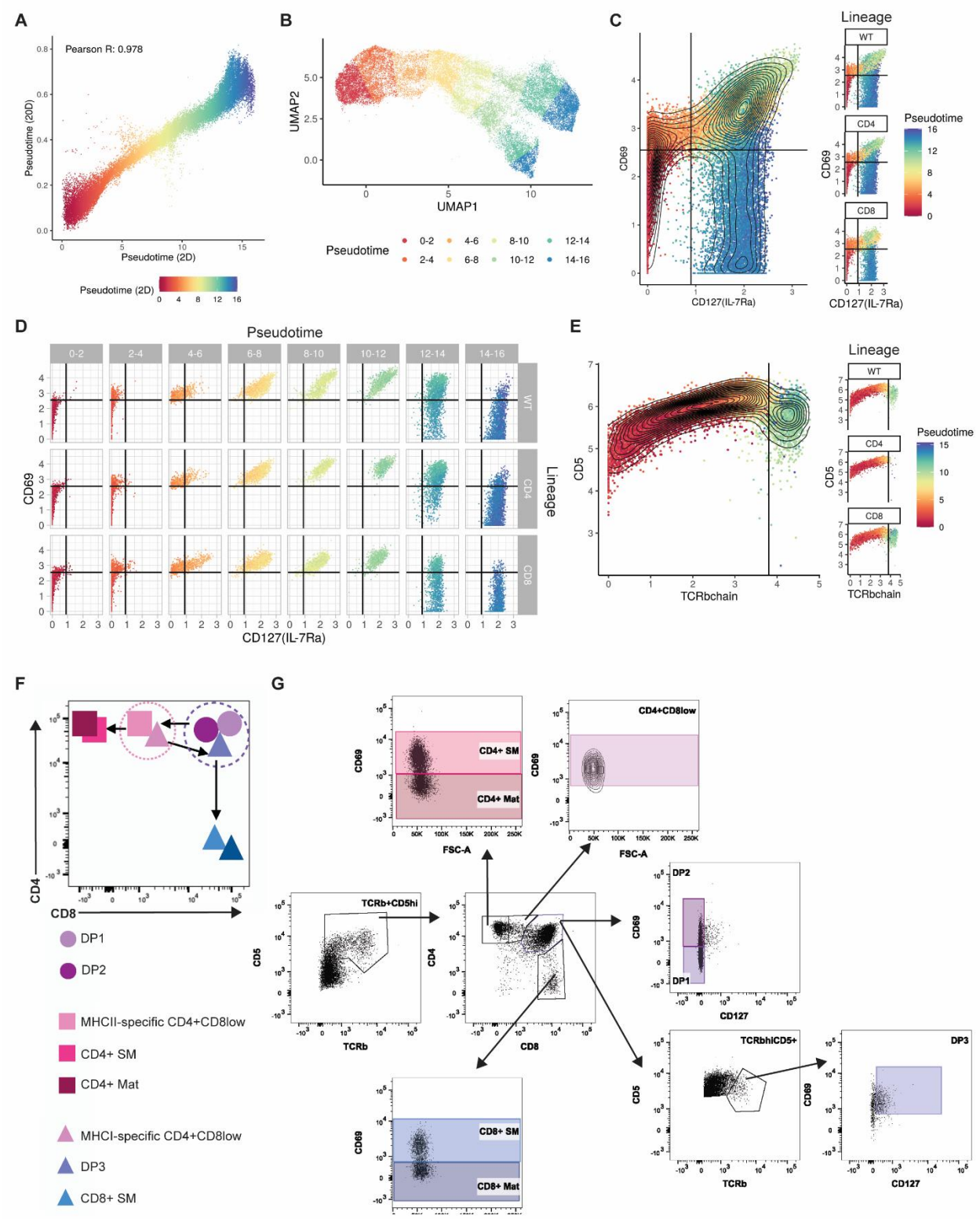


Figure S2: Pseudotime inference identifies intermediate thymocyte stages that can be isolated by FACS.
(caption on following page)

Figure S2: Pseudotime inference identifies intermediate thymocyte stages that can be isolated by FACS. (A) Correlation between Slingshot pseudotime inferred from the full 20-dimensional totalVI latent space and a 2-dimensional UMAP projection of the 20-dimensional latent space. **(B)** UMAP plot of the totalVI latent space from positively-selected thymocytes. Cells are colored according to placement in one of eight bins uniformly spaced over 2D pseudotime for visualization. **(C)** In silico flow cytometry plots of log(totalVI denoised expression) of CD127(IL-7Ra) and CD69 from positively-selected thymocytes (left) and the same cells separated by lineage (right). Cells are colored by pseudotime. **(D)** In silico flow cytometry plot of data as in (C) separated by lineage and pseudotime. **(E)** In silico flow cytometry plots of log(totalVI denoised expression) of TCRbchain and CD5 from DP thymocytes (left) and the same cells separated by lineage (right). Cells are colored by pseudotime. Among DP thymocytes, the DP3 population is TCRbchain high, CD127+, and CD69+. **(F)** Schematic of a CD4 vs CD8 biaxial plot to identify gated populations in adult thymocytes. Cells were gated into 8 subsets: double positive 1 (DP1), double positive 2 (DP2), double positive 3 (DP3), MHCII-specific CD4+CD8lo (CD4+CD8lo), CD4+ semimature (CD4+ SM), CD4+ mature (CD4+ Mat), MHCI-specific CD4+CD8lo (CD4+CD8lo), CD8+ semimature (CD8+ SM) and CD8+ mature (CD8+ Mat). Circles represent lineage uncommitted cells, squares represent CD4 lineage committed, and triangles represent CD8 lineage committed cells. **(G)** Representative FACS gating strategy for thymocyte populations in adult mice. Thymocytes were harvested from 6-8-week-old WT (B6), β2M-/- or MHCII-/- mice. Cell populations were gated to include lymphocytes, exclude forward scatter and side scatter doublets, include live cells, include TCRβ+CD5int/hi, then on CD4 vs CD8. Cell populations were gated into the following subsets based upon cell surface marker expression: double positive 1 (DP1; CD4+CD8+CD127-CD69-), double positive 2 (DP2; CD4+CD8+CD127-CD69+), double positive 3 (DP3; TCRβhiCD5+CD127+CD69+), MHCII-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loCD69+, only in β2M-/- mice), CD4+ semimature (CD4+ SM; CD4+CD8-CD69+), CD4+ mature (CD4+ Mat; CD4+CD8-CD69-), MHCI-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loCD69+, only in MHCII-/- mice), CD8+ semimature (CD8+ SM; CD8+CD69+) and CD8+ mature (CD8+ Mat; CD8+CD69-).

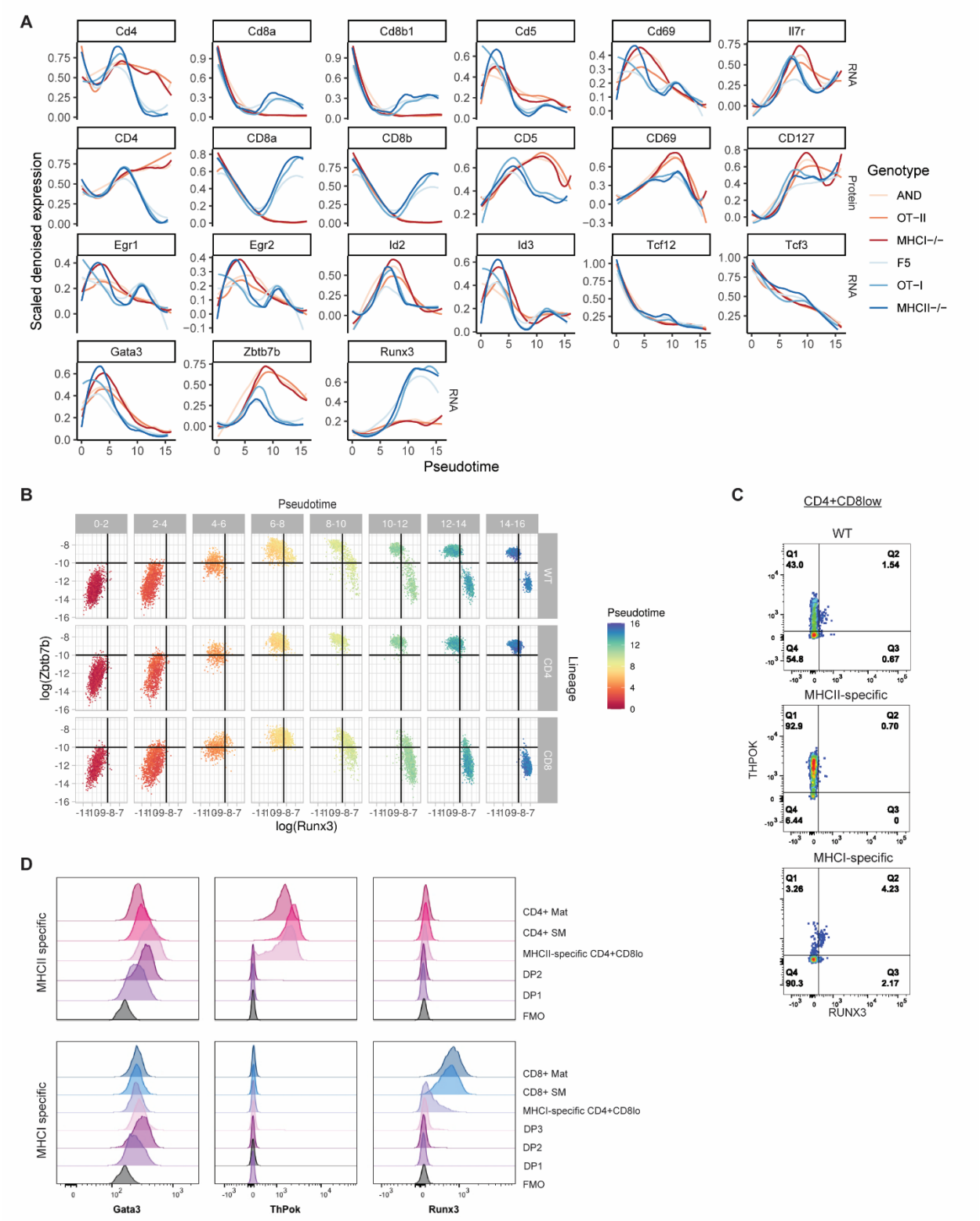


Figure S3: CITE-seq and fluorescence-based flow cytometry reveal the timing of expression for transcription factors and other features of CD4/CD8 development. (caption on following page)

Figure S3: CITE-seq and fluorescence-based flow cytometry reveal the timing of expression for transcription factors and other features of CD4/CD8 development. (A) Expression of RNA and protein features over pseudotime by genotype. Features are totalVI denoised expression values scaled per feature and smoothed by loess curves. **(B)** In silico flow cytometry plots of log(totalVI denoised expression) of *Runx3* and *Zbtb7b* from positively-selected thymocytes separated by pseudotime. **(C)** Representative FACS dot plots displaying RUNX3 vs THPOK protein expression in CD4+CD8lo (CD4+CD8loCD69+) cells from WT, MHCII-specific (β2M-/-), and MHCI-specific (MHCII-/-) adult mice. Positive staining and gates were determined using fluorescence minus one (FMO) controls. **(D)** Transcription factor protein expression in adult thymocyte populations. Representative histograms displaying GATA3, THPOK, and RUNX3 transcription factor expression detected by FACS intracellular staining in MHCII-specific (β2M-/-) and MHCI-specific (MHCII-/-) thymocyte populations. Thymocyte populations were gated on lymphocytes, excluding forward scatter and side scatter doublets, live cells, TCRβ+CD5int/hi then on CD4 vs CD8. Cell populations were gated into the following subsets based upon cell surface marker expression: double positive 1 (DP1; CD4+CD8+CD127-CD69-), double positive 2 (DP2; CD4+CD8+CD127-CD69+), double positive 3 (DP3; CD4+CD8+TCRβhiCD5+CD127+CD69+), MHCII-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loCD69+, only in β2M-/- mice), CD4+ semimature (CD4+ SM; CD4+CD8-CD69+), CD4+ mature (CD4+ Mat; CD4+CD8-CD69-), MHCI-specific CD4+CD8lo (CD4+CD8lo; CD4+CD8loCD69+, only in MHCII-/- mice), CD8+ semimature (CD8+ SM; CD8+CD69+) and CD8+ mature (CD8+ Mat; CD8+CD69-). Positive staining was determined using a fluorescence minus one (FMO) control.

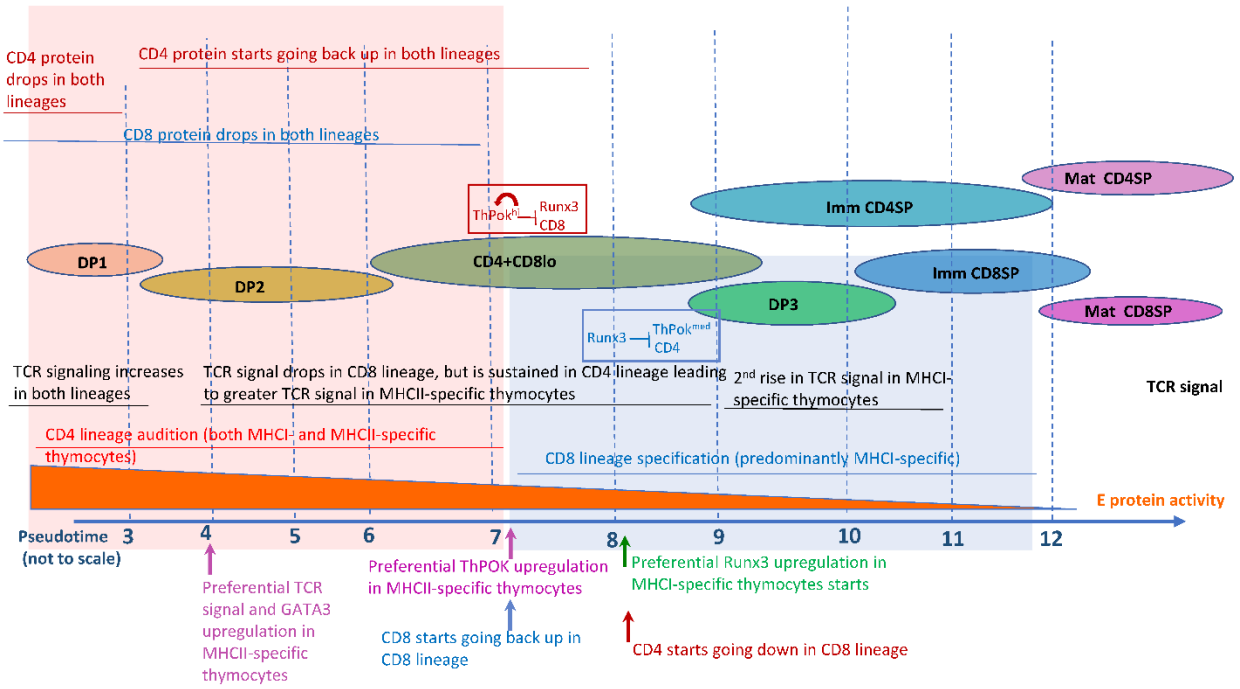


Figure S4: A temporal model for CD4 versus CD8 T cell lineage commitment. Key changes in mRNA and protein inferred from CITE-seq data are displayed from left to right in their order of occurrence based on pseudotime. Colored ovals indicate the relative order of key thymocyte stages as defined by cell surface markers. Black text and lines indicate dynamic changes in TCR signaling. Shaded red area indicates the time window during which both MHCI and MHCII specific thymocytes audition for the CD4 fate, corresponding to upregulation of GATA3 followed by THPOK. Shaded blue area indicates the later time window during which thymocytes that fail the CD4 audition (mostly MHCI specific) receive CD8 lineage reinforcement and survival signals. Bold upward pointing arrows at bottom indicate relative timing of lineage defining gene expression changes. During the earlier CD4 auditioning phase of positive selection (red shading), most MHCII-specific thymocytes receive moderate, persistent TCR signals, allowing them to lock in the CD4 fate (red rectangle) by fully upregulating THPOK, activating the THPOK positive autoregulation loop (red circular arrow, (Muroi et al., 2008)), leading to repression of CD8 and RUNX3. In contrast most MHCI-specific thymocytes receive weaker more transient signals during this phase (due to the combined effect of weaker LCK recruitment by the CD8 co-receptor, drop in CD8 surface expression and increase in the negative regulator CD5 (Chan et al., 1999)). During the later CD8 lineage specification window (blue shading), the CD8 SP enhancer and RUNX3 are activated, likely due to the maturation-associated drop in E protein activity (orange triangle, (Jones-Mason et al., 2012)). Continued upregulation of RUNX3 represses both THPOK and CD4, while further enhancing CD8 expression. In addition, an increase in TCR sensitivity during the CD8 lineage specification phase (due to the drop in negative regulator CD5, rise in ZAP70, and increase in ion channel components (e.g., KCNA2 and TMIE (Lutes et al 2021)) leads to a second wave of TCR signaling. TCR signaling at this late stage of CD8 specification may provide survival signals and well as further upregulating RUNX3, and would serve to ensure the elimination of any MHCII-specific thymocytes that failed the CD4 audition phase.

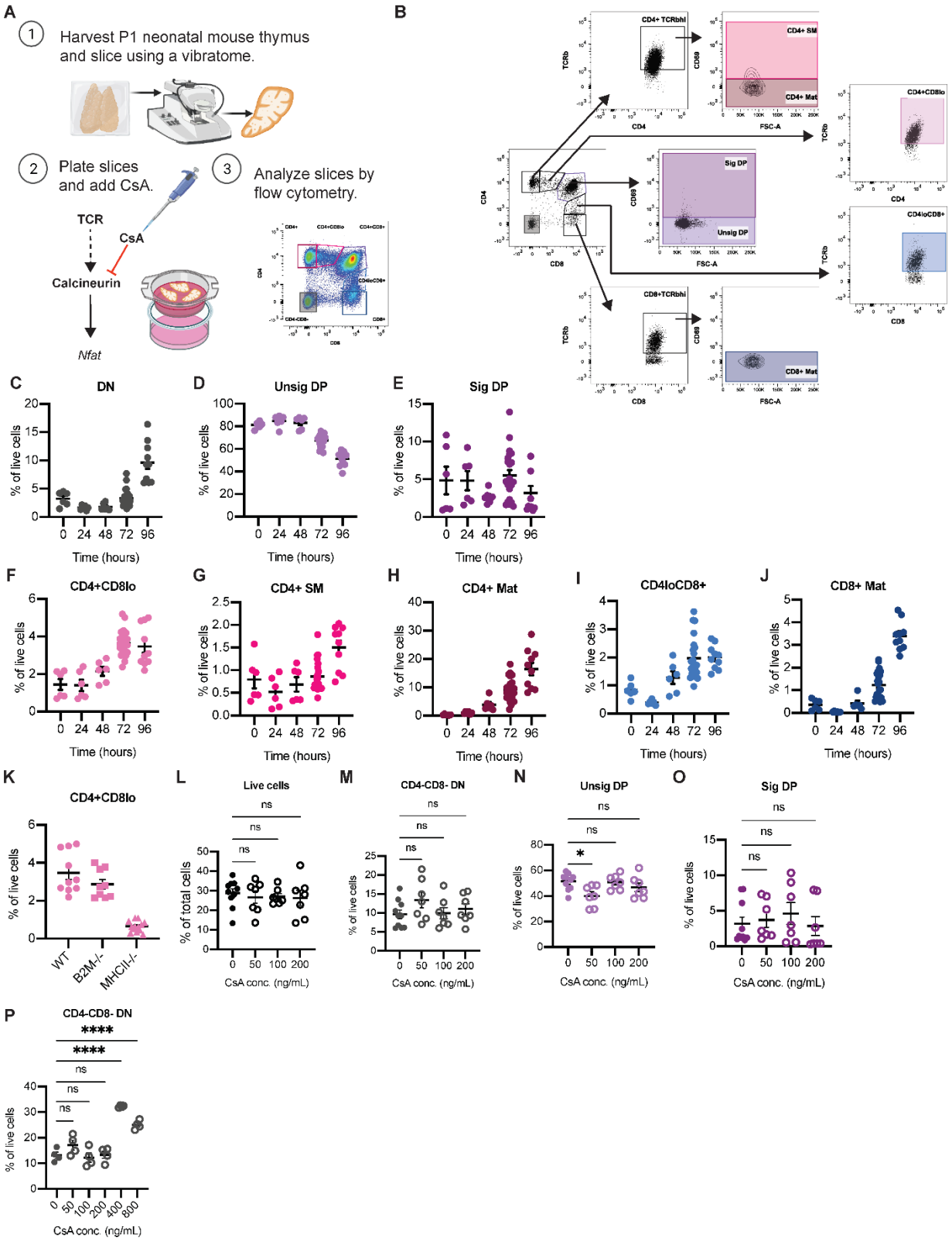


Figure S5: Cell populations in neonatal thymic slice cultures with Cyclosporin A. (caption on following page)

Figure S5: Cell populations in neonatal thymic slice cultures with Cyclosporin A. Thymic slices were isolated from postnatal day 1 (P1) mice and cultured at time point 0 in media alone containing no cyclosporin A (CsA) or with various concentrations of CsA for up to 96 hours. Thymic slices were collected and analyzed at indicated time points via flow cytometry to quantify cell populations. **(A)** Experimental overview of neonatal thymic slice cultures. Illustrations were created using Biorender.com. **(B)** Representative neonatal slice culture FACS gating strategy. Cell populations were gated to include lymphocytes, exclude forward scatter and side scatter doublets, include live cells, then on CD4 vs CD8. Thymocytes were further gated into double negative (DN; CD4-CD8-), unsignaled double positive (Unsig DP; CD4+CD8+CD69-), signaled double positive (Sig DP; CD4+CD8+CD69+), CD4+CD8lo (CD4+CD8loTCRb+), CD4+ semimature (CD4+ SM; CD4+CD8-TCRbhiCD69+), CD4+ mature (CD4+ Mat; CD4+CD8-TCRbhiCD69-), CD4loCD8+ (CD4loCD8+TCRb+), CD8+ mature (CD8+ Mat; CD4-CD8+TCRbhiCD69-). **(C-J)** WT neonatal slice time course experiments. Data is compiled from 8 independent experiments, where each symbol represents a thymic slice. Plots display the frequency of **(C)** CD4-CD8- DN, **(D)** unsig DP, **(E)** sig DP, **(F)** CD4+CD8lo, **(G)** CD4+ SM, **(H)** CD4+ Mat, **(I)** CD4loCD8+, **(J)** CD8+ Mat cells in thymic slices after 0, 24, 48, 72 and 96 hours of culture. **(K)** Frequency of CD4+CD8lo cells in WT, MHCII-specific (B2M-/-), and MHCI-specific (MHCII-/-) neonatal slices after 96 hours in culture. All genotypes possess a CD4+CD8lo cell population. Each symbol represents a thymic slice. **(L-O)** Addition of CsA to WT B6 P1 thymic slice cultures for 96 hours does not affect the **(L)** frequency (% of total) of live cells, **(M)** frequency (% of live) of CD4-CD8- DN cells, **(N)** of unsig DP or **(O)** of sig DP. Data is compiled from 3 independent experiments. where each symbol represents a thymic slice. **(P)** Addition of high concentrations (400 and 800 ng/mL) of CsA increases frequency of CD4-CD8- DN cells. Data is representative of 2 independent experiments, n= 4 for each concentration of CsA, where each symbol represents a thymic slice. Data in (L-P) were analyzed using an ordinary one-way ANOVA. NS is not significant, *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

Chapter 5

Concluding remarks

Single-cell multi-omics analysis is a quickly developing field on both the experimental and computational fronts. These approaches have great potential to expand biological knowledge of cell types and the key molecules that define them. Multi-omics facilitates the drawing of connections between multiple views of cell identity and enables connecting the molecules themselves into regulatory pathways, signal transduction cascades, and a more detailed understanding of how cellular mechanisms are carried out. I have thus far described new approaches to leverage multi-omic measurements by performing joint analyses that combine RNA and protein data. I have applied these approaches to illuminate the fundamental biological processes of cell development and lineage commitment within the context of the thymus. Here, I briefly discuss promising directions for the previously discussed work.

The work presented in Chapter 4 characterizes thymocytes that are developing into CD4 and CD8 T cells on the basis of their transcriptomes and surface proteins from a snapshot in time of this continuous developmental trajectory. The measurement of additional modalities would be a natural next step to gain a more complete picture of this process. Future work would benefit from connecting these multi-omic profiles with the spatial locations of cells within the thymus, as the physical movement of cells from the thymic cortex to the medulla plays a role in maturation. Beyond providing spatial locations, imaging studies could more precisely quantify the timing of developmental events, identify the role of cell-cell interactions in thymic migration and developmental progression, and more directly track TCR signaling events using reporter systems. The measurement of chromatin accessibility or transcription factor binding could additionally inform the regulatory networks governing commitment to and enforcement of the CD4 and CD8 lineages. While the connection between TCR sequence and antigen-MHC specificity remains an open question, the additional measurements of TCR sequences could aid our understanding of how a thymocyte with a given TCR sequence ultimately becomes specified towards the CD4 or CD8 lineage.

This work primarily focused on the lineage commitment between CD4 and CD8 T cells. However, the methods described here could be readily applied to study the development of other cell types. For instance, the dataset presented in Chapter 4 contains multiple thymic subsets of interest, such as regulatory T cells (Tregs) that play an important role in autoimmune disease. Moreover, the analytical approaches applied here could provide a useful approach to investigate cells outside of the thymus, such as the differentiation of T cells in the periphery in response to infection as well as the development of non-immune cell types.

As described in Chapter 2, while the totalVI model was designed specifically for the analysis of paired transcriptome and surface protein data, the underlying conceptual framework is readily extensible to include additional measurements. Given that multiple measurements are generated from the same cell, a future model could jointly analyze RNA and protein data along with modalities such as chromatin accessibility or spatial locations. As with the totalVI model, the

addition of each new modality would require careful consideration of the technical nuances of the measurement. The analysis of multiple molecule types from the same cell (e.g., DNA, RNA, protein) invites future research on the fundamental dynamics of gene expression including the kinetics of transcription and translation. It also presents the opportunity to reconstruct gene regulatory networks by linking multiple molecule types, such as connecting a protein signaling event at the cell surface with downstream transcriptional events, or transcriptional regulation with subsequent surface protein expression changes. Ongoing work has extended the totalVI model to correct continuous covariates and allow for mapping new data onto an already trained model for a reference dataset, speeding dataset integration for large-scale atlas projects. In future work, the paired analysis of RNA and proteins with totalVI (potentially along with additional modalities) could potentially improve automated cell type annotation by relying on all available features to characterize cell types and cell states.

As a tool for multi-omic data analysis, totalVI can be even more useful to the scientific community if it is made accessible and easy to use for diverse applications. totalVI is freely available through the scvi-tools package (scvi-tools.org), including extensive tutorials and documentation to increase user accessibility. In addition to streamlining the code to improve usability, ongoing work on the scvi-tools package aims to make the models themselves more modular and easily adaptable by developers for the creation of new models. Because multi-omics experiments and computational analyses can be complicated and often involve tricks of the trade learned by experience that are often unpublished, there can be a high barrier for researchers to attempt to use these methods. Chapter 3 translates some of the theory behind totalVI into an intuition for multi-omics analysis and practical guidance for performing multi-omics experiments and joint analysis. In enumerating many of the numerous experimental and computational variables and offering recommendations for navigating these decision points, I hope to enable other researchers to apply these methods to new questions and biological systems.

Appendix 1 - Supplementary Information for Joint probabilistic modeling of single-cell multi-omic data with totalVI

Contents

Supplementary Tables

Dataset name	Antibody panel	Day	Mouse	Tissue	Cells (captured)	Cells (post-filtering)
SLN111-D1	111	1	A	Spleen; Lymph Node	11,160	9,264
SLN111-D2	111	2	B	Spleen; Lymph Node	9,017	7,564
SLN208-D1	208	1	A	Spleen; Lymph Node	10,777	8,715
SLN208-D2	208	2	B	Spleen; Lymph Node	8,921	7,105

Supplementary Table 1: Summary of spleen and lymph node datasets. Each dataset was processed in a separate 10x lane. Each day indicates a 10x run. Cells captured is the number of cells reported by Cell Ranger.

Cell type	Protein	totalVI ROC AUC	GMM ROC AUC
B cells	CD19	**0.9998**	0.9997
B cells	CD45R-B220	**0.9999**	0.9974
B cells	CD20	**0.9995**	0.8600
B cells	I-A-I-E	**0.9941**	0.9725
T cells	CD5	**0.9998**	0.9974
T cells	TCRbchain	**0.9999**	0.9976
T cells	CD90.2	**0.9999**	0.9986
T cells	CD28	**0.9999**	0.8328
CD4 T cells	CD4	**0.9999**	0.9969
CD8 T cells	CD8a	**0.9999**	0.9985
CD8 T cells	CD8b	**0.9998**	0.9980

Supplementary Table 2: Classification of cell types by marker proteins (SLN111-D1 dataset.) Performance of totalVI and a Gaussian mixture model (GMM) fit on all cells for each protein of the SLN111-D1 dataset. Area under the receiver operating characteristic curve (ROC AUC score) was calculated using as input either the totalVI foreground probability or GMM foreground probability where the indicated cell type was the positive population out of all B and T cells. AUC results were truncated at four decimal places. Bolded ROC AUC scores indicate higher values (better performance). Highlighted in blue are two proteins for which totalVI noticeably outperformed the GMM.

Cell type	Protein	Isotype Control	totalVI ROC AUC	GMM ROC AUC	GMM norm1 ROC AUC	GMM norm2 ROC AUC
B cells	CD19	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9996	0.9988	0.9988
B cells	CD45R-B220	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9973	0.9959	0.9961
B cells	CD20	IsotypeCtrlRatIgG2b_k	**0.9999**	0.7197	0.7528	0.7494
B cells	I-A-I-E	IsotypeCtrlRatIgG2b_k	**0.9984**	0.9783	0.9695	0.9695
T cells	CD5	IsotypeCtrlRatIgG2a_k	**0.9998**	0.9966	0.9938	0.9938
T cells	TCRbchain	IsotypeCtrlArm.HamsterIgG	**0.9999**	0.9943	0.7725	0.9762
T cells	CD90.2	IsotypeCtrlRatIgG2b_k	**0.9999**	0.9988	0.8316	0.9984
T cells	CD28	NA	**0.9997**	0.7805	0.7805	0.7805
CD4 T cells	CD4	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9966	0.9966	0.9966
CD8 T cells	CD8a	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9991	0.7952	0.9989
CD8 T cells	CD8b	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9993	0.8606	0.9989

Supplementary Table 3: Classification of cell types by marker proteins (SLN208-D1 dataset). Performance of totalVI and a GMM fit on all cells for each protein of the SLN208-D1 dataset as in Supplementary Table 2. GMM norm1 indicates that data were normalized using an isotype control as in Cumulus [1] prior to fitting the GMM; GMM norm2 indicates that data were normalized using a modification to the Cumulus method (Methods). Bolded ROC AUC scores indicate highest value for each protein. Notable result is highlighted in blue.

Cell type annotation	Selected markers	Selected references
Activated CD4 T cells	*Itm2a, Cd69*	[2]
B1 B cells	*Bhlhe41*, CD43, CD19	[3, 4]
CD122+ CD8 T cells	CD122, CD62L, CD183(CXCR3), CD8a	[5, 6]
CD4 T cells	CD4	[7]
CD8 T cells	CD8a, CD8b	[7]
cDC1s	*Clec9a, Cd8a, Xcr1*, CD11c	[8, 9]
cDC2s	*Itgax, Cd4*, CD11c	[8, 9]
Cycling B/T cells	*Birc5, Top2a, Mki67*	[10]
Erythrocytes	*Hbb-bs, Hbb-bt*	[11]
GD T cells	*Cd3e, Tcrg-c1, Tcrg-c2, Maf, Il17re*	[12]
ICOS-high Tregs	*Foxp3*, CD4, ICOS	[13, 14]
Ifit3-high B cells	*Ifit3*, CD19	
Ifit3-high CD4 T cells	*Ifit3*, CD4	
Ifit3-high CD8 T cells	*Ifit3*, CD8a	
Ly6-high monocytes	*Ly6c2, Fn1, F13a1*	[15]
Ly6-low monocytes	*Cd36, Cd300e, Fabp4*	[16]
Mature B cells	IgD, CD23, CD19	[17]
Migratory DCs	*Slco5a1, Anxa3, Nudt17, Adcy6, Cacnb3*	[18]
MZ B cells	CD21, CD19	[17]
MZ/Marco-high macrophages	*Cd209b, Marco*	[19]
Neutrophils	*S100a8*	[20]
NK cells	NK-1.1, *Gzma, Ncr1*	[7, 21]
NKT cells	NK-1.1, *Cd3e, Ccl5, Klrd1*	[7, 22]
pDCs	*Siglech, Irf8, Runx2*, CD11c	[8, 23, 24]
Plasma B cells	*Jchain*	[25]
Red-pulp macrophages	F4-80, *C1qa, C1qb, Hmox1, Vcam1*	[26]
Transitional B cells	CD93, CD24, CD19	[27, 28]
Tregs	*Foxp3*, CD4, CD357(GITR)	[29]

Supplementary Table 4: Annotated cell types and selected markers in the spleen and lymph node datasets. cDC1: Conventional dendritic cell 1. cDC2: Conventional dendritic cell 2. GD: Gamma/delta. MZ: Marginal zone. NK: Natural killer. NKT: Natural killer T. pDC: Plasmacytoid dendritic cell. Treg: Regulatory CD4 T cell.

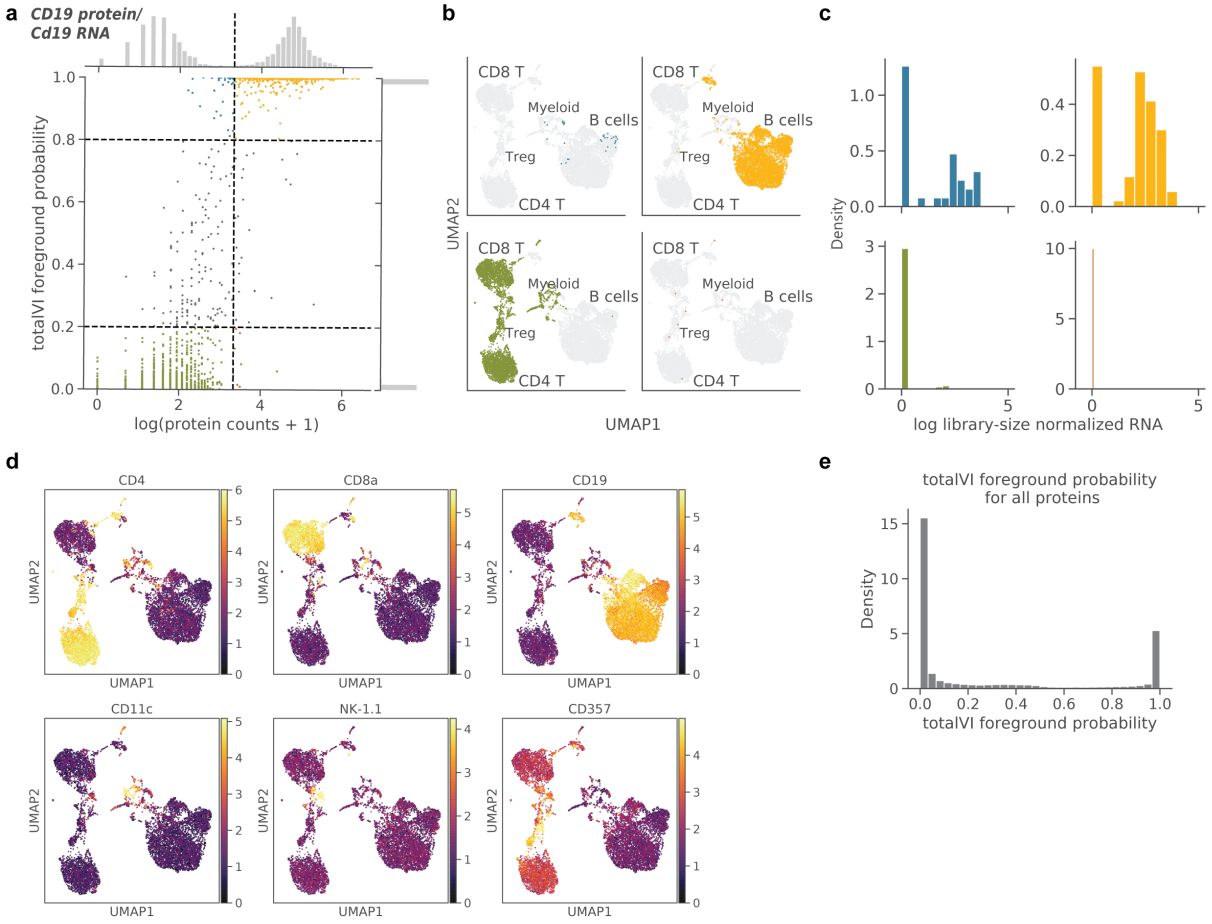
Name	RNA reads	Protein reads	RNA UMI	Protein UMI
SLN111-D1	34,717	4,733	4,392	2,785
SLN111-D2	45,765	6,542	2,121	3,419
SLN208-D1	33,569	5,513	4,561	2,956
SLN208-D2	43,821	3,961	2,102	2,261

Supplementary Table 5: Sequencing statistics for spleen and lymph node datasets. Sequencing statistics calculated per 10x lane by Cell Ranger. RNA reads: mean reads per cell from RNA. Protein reads: mean reads per cell from antibody barcodes. RNA UMI: median UMI counts per cell from RNA. Protein UMI: median UMI counts per cell from antibody barcodes.

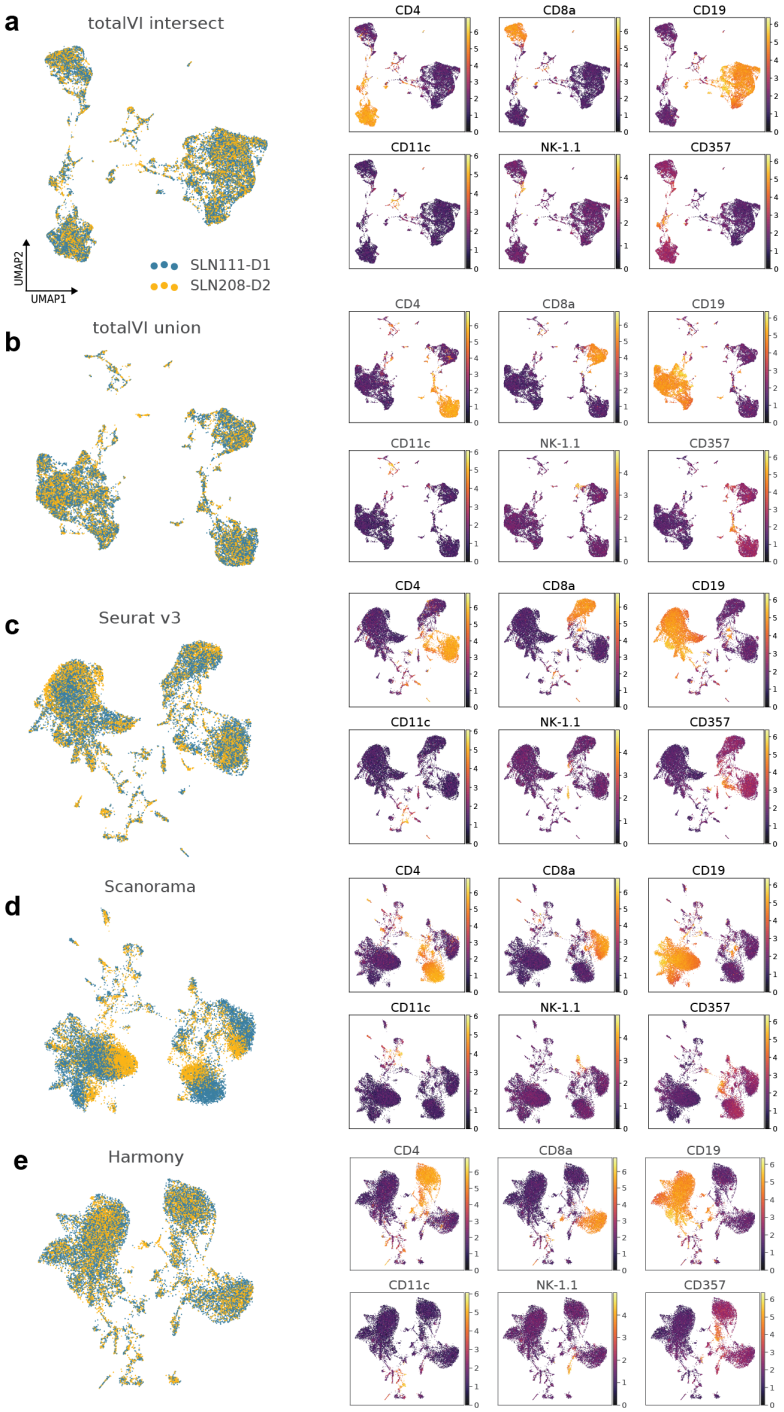
Dataset	No. cells	Pct. Mito	Protein Lib Size Range	No. Genes Expr.	RNA lib size
PBMCK10k	6,855	$< 10\%$	[400, 20,000]	$<$ 4,500	$<$ 20,000
PBMC5k	3,994	$< 20\%$	[400, 20,000]	$<$ 4,500	$<$ 20,000
MALT	6,838	$< 15\%$	[400, 20,000]	$<$ 5,000	$<$ 30,000

Supplementary Table 6: Summary of filtering parameters for publicly available datasets. Ranges indicate criteria for retained cells.

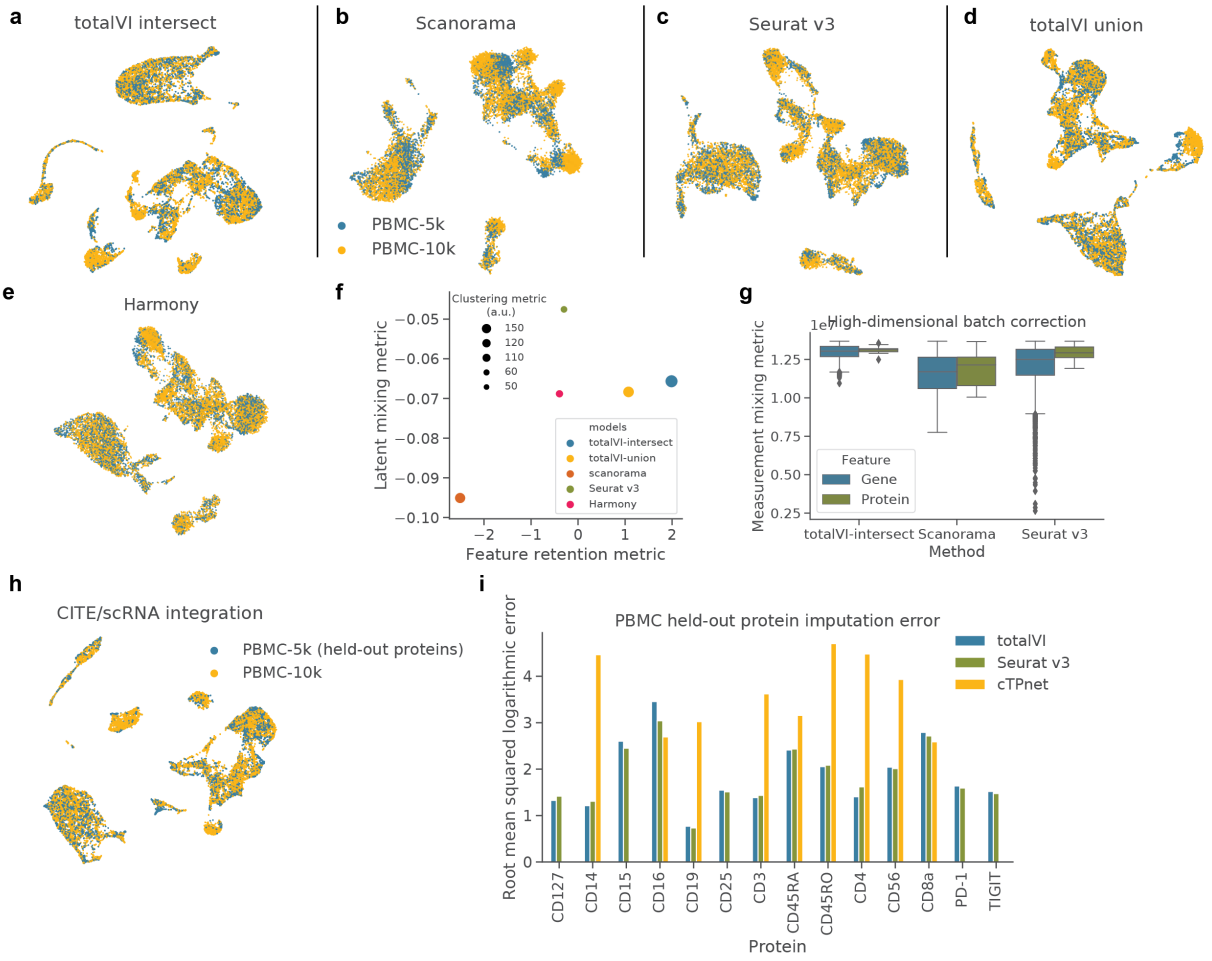
Supplementary Figures



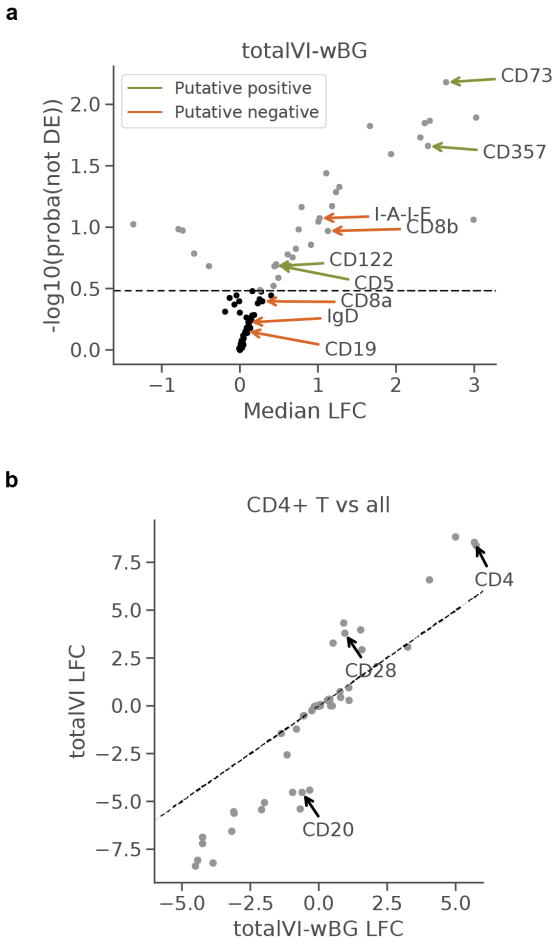
Supplementary Figure 1: totalVI decouples protein foreground and background. totalVI was applied to the SLN111-D1 dataset. **a-c**, CD19 protein (encoded by *Cd19* RNA). **(a)** totalVI foreground probability vs $\log(\text{protein counts} + 1)$. Vertical line denotes protein foreground/background cutoff determined by a GMM. Horizontal lines denote foreground probability of 0.2 and 0.8. Cells with foreground probability greater than 0.8 or less than 0.2 are colored by quadrant, while the remaining cells are gray. **(b)** UMAP plots of the totalVI latent space. Each quadrant contains cells from the corresponding quadrant of (a) in color with the remaining cells in gray. **(c)** RNA expression (log library-size normalized; Methods) for cells colored in (a). **d**, UMAP plots of the totalVI latent space colored by (log(counts + 1)) of cell type marker proteins (Supplementary Table 4). **e**, totalVI foreground probability for all proteins across all cells in the SLN111-D1 dataset.



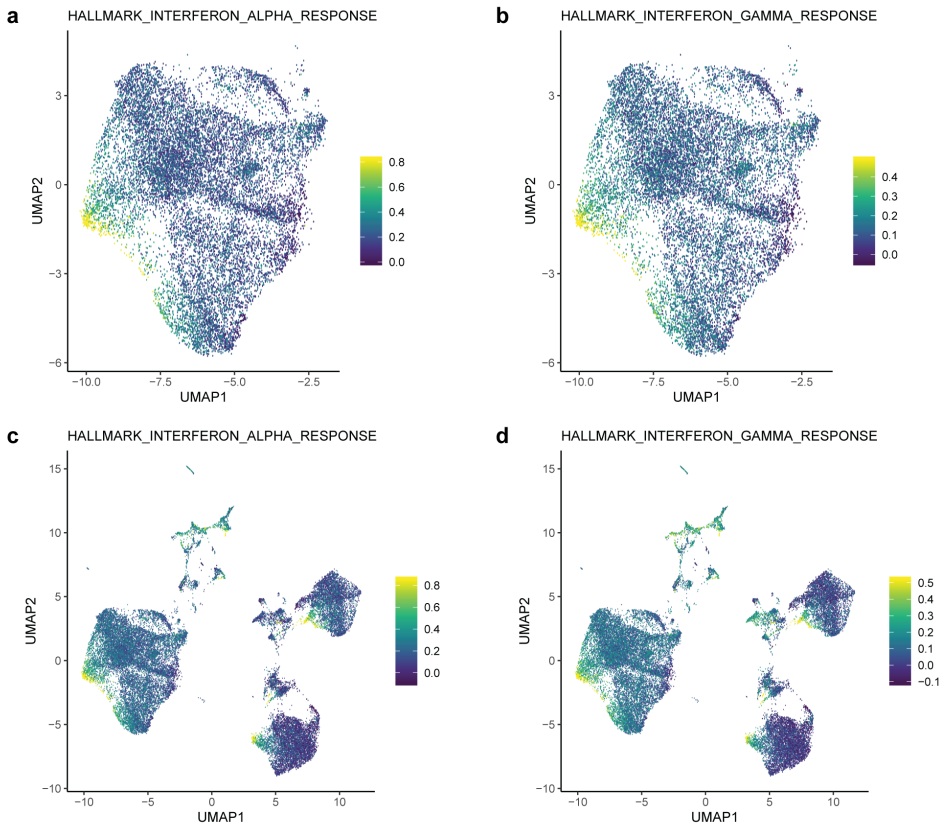
Supplementary Figure 2: UMAP embeddings of integration methods on spleen and lymph node data. **a-e**, For each method, UMAP plots colored by dataset, and by $\log(\text{counts} + 1)$ of key marker proteins (Supplementary Table 4).



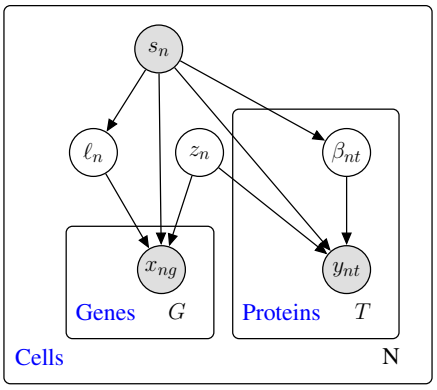
Supplementary Figure 3: Benchmarking of integration methods on PBMC data. Integration methods were applied to PBMC10k and PBMC5k. **a-e**, UMAP plots of integrated latent spaces. **f**, Latent mixing metric, feature retention metric, and clustering metric for each method (Methods). **g**, Measurement mixing metric applied individually to each batch-corrected feature (computed for $n = 4000$ genes and $n = 14$ proteins). Box plots indicate the median (center lines), interquartile range (hinges), whiskers at 1.5x interquartile range. **h**, UMAP plot of integrated latent space from totalVI union mode when holding out the proteins from PBMC5k. **i**, Root mean squared logarithmic error between imputed and observed proteins from PBMC5k for totalVI and Seurat v3. cTP-net did not provide predictions for CD127, CD15, CD25, PD-1, or TIGIT.



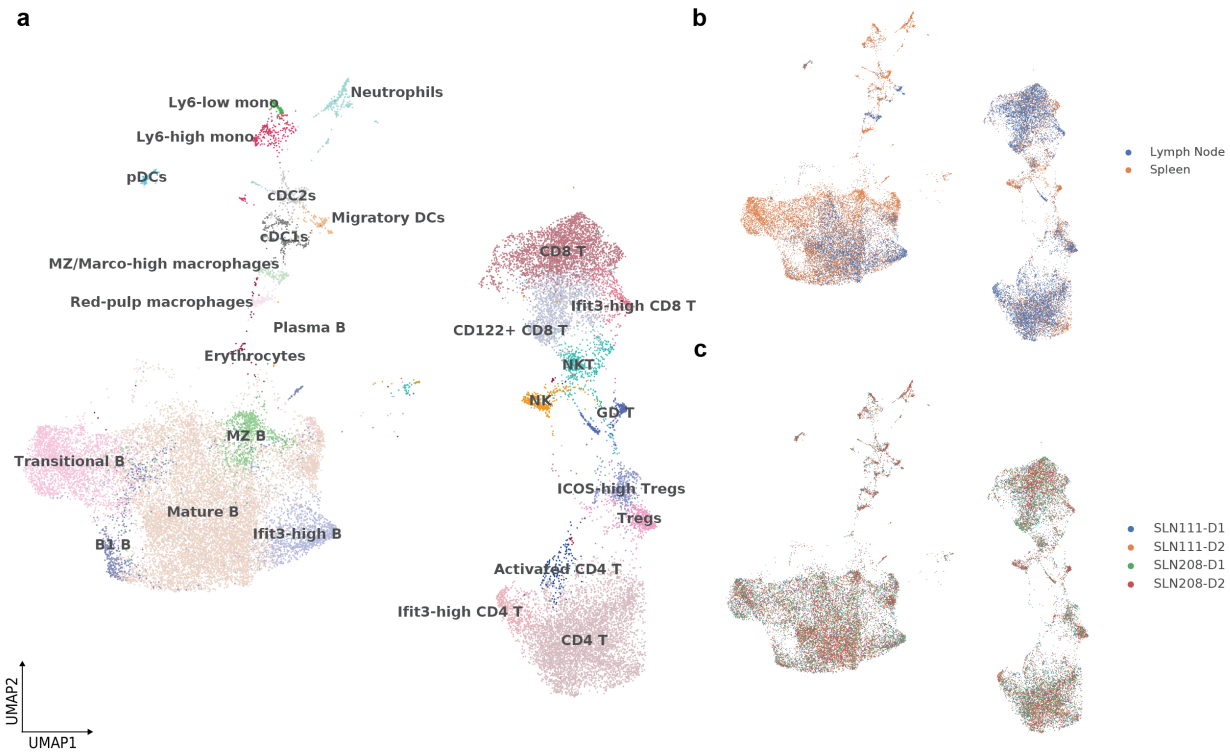
Supplementary Figure 4: Differential expression using totalVI-wBG, which does not correct for the protein background component. **a** Volcano plot for the ICOS-high Tregs vs CD4 T cells test. Putative positives and negatives are highlighted (Methods). **b** The LFC estimates for totalVI and totalVI-wBG on the CD4 T cells versus all others test.



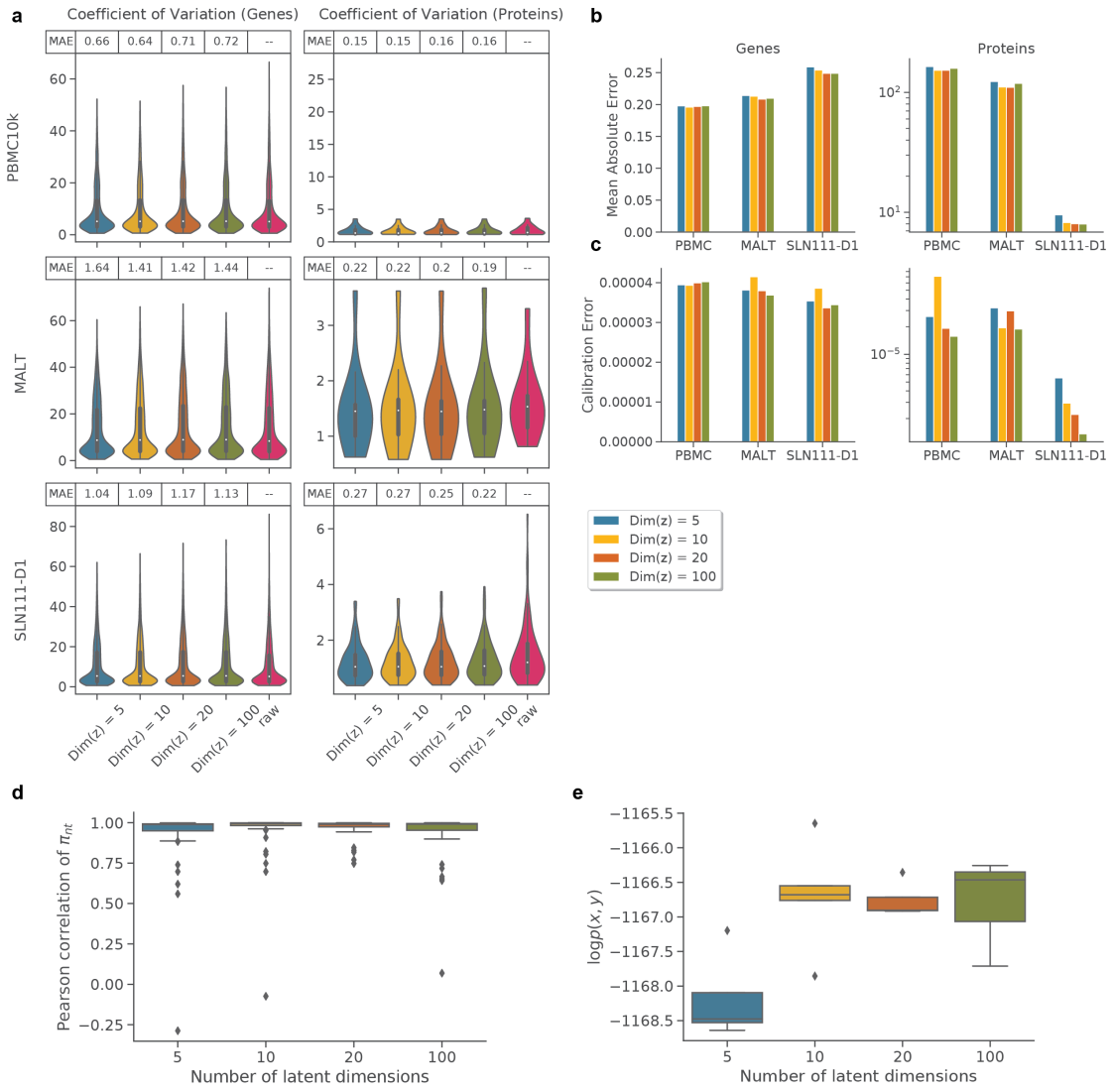
Supplementary Figure 5: Interferon signatures in the mouse spleen and lymph node. a, b, UMAP of totalVI latent space for B cells of the SLN-all dataset **(a)** colored by the Hallmark Interferon Alpha Response signature score and **(b)** colored by the Hallmark Interferon Gamma Response signature score (Methods). **c, d**, Same as in (a, b), but for all cells in SLN-all.



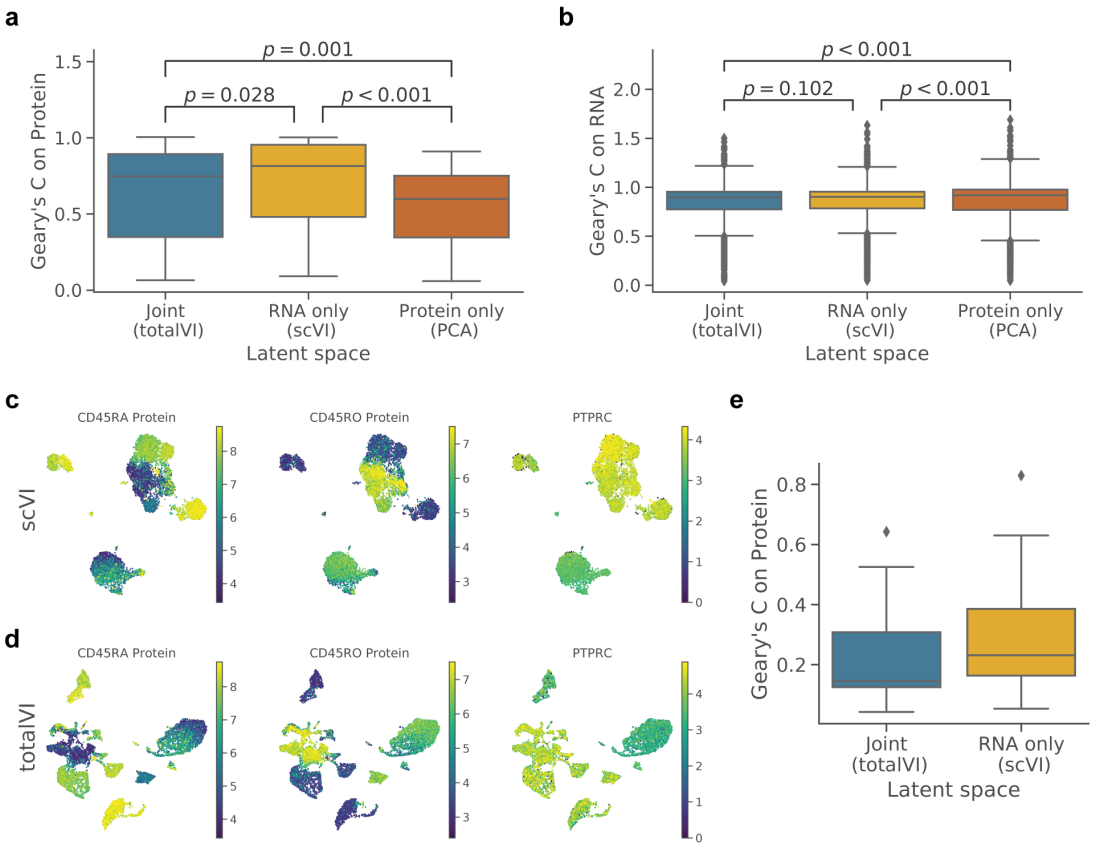
Supplementary Figure 6: totalVI probabilistic graphical model. Shaded nodes represent observed random variables. Unshaded nodes represent latent variables. Edges denote conditional independence. Rectangles ("plates") represent independent replication.



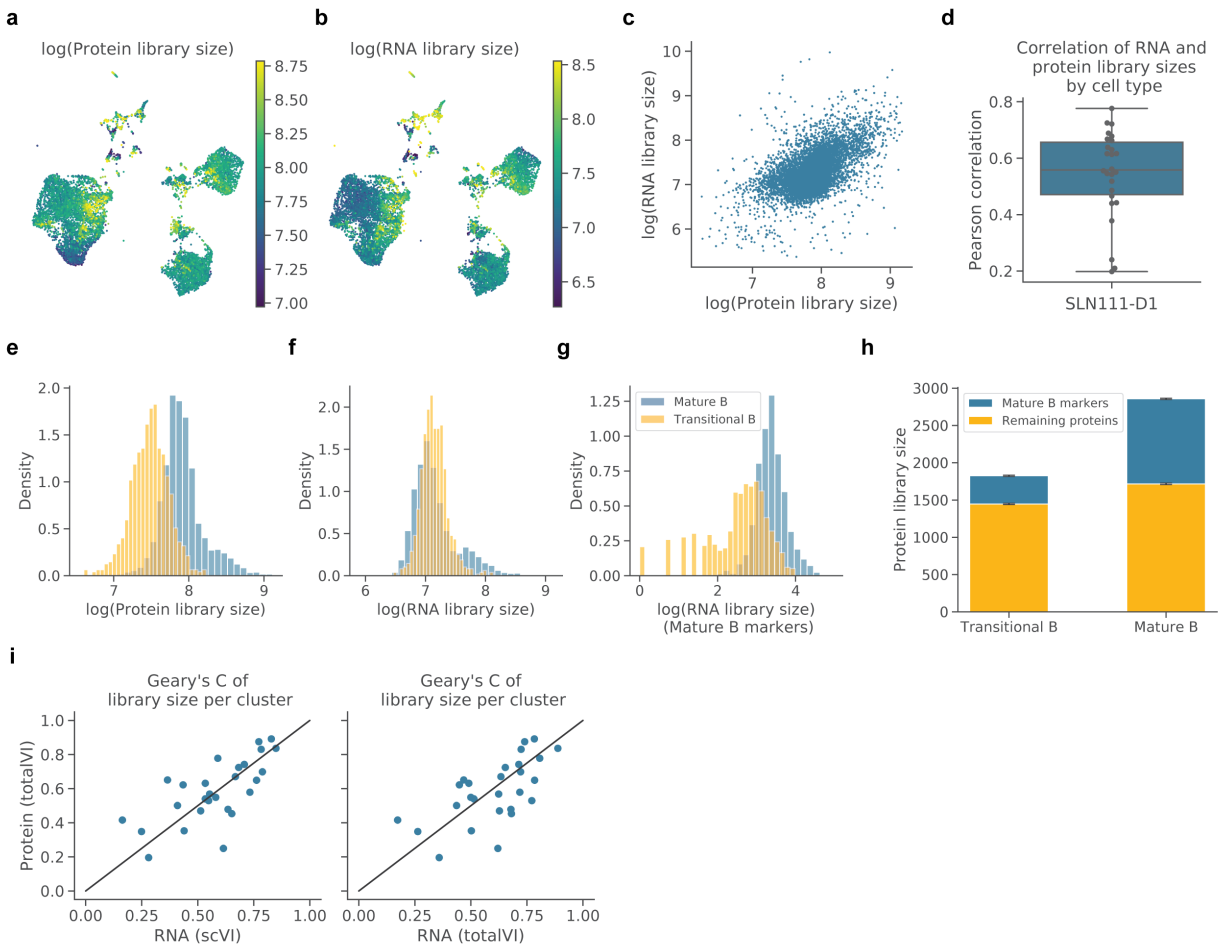
Supplementary Figure 7: Integration of SLN-all with union of panels. **a-c**, UMAP plot of SLN-all colored by **(a)** cell types derived from manual annotation of model run with intersection of panels, **(b)** tissue, and **(c)**, dataset.



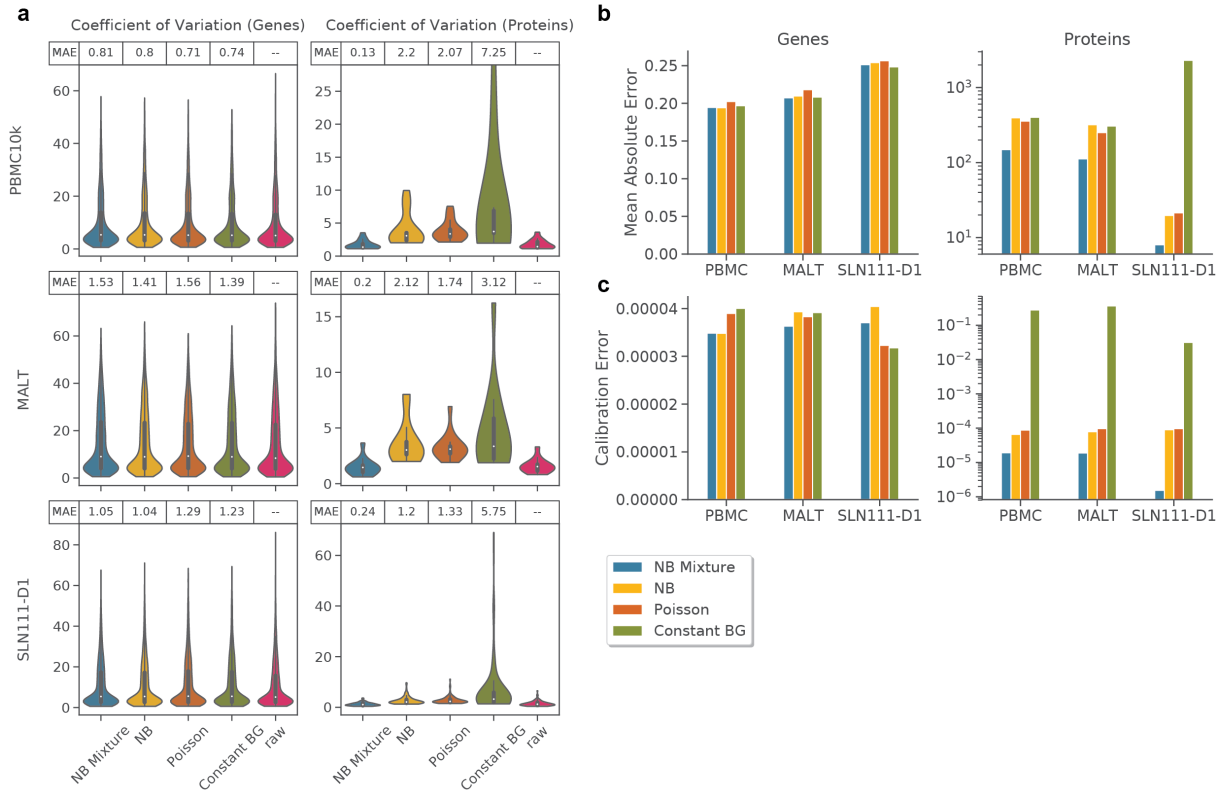
Supplementary Figure 8: Evaluation of totalVI across choice of number of latent dimensions. **a**, Posterior predictive check of coefficient of variation (CV) of genes and proteins. For totalVI with 5, 10, 20, and 100 latent dimensions, the average CV from posterior predictive samples was computed for each feature. Violin plots summarize the distribution of CVs for genes and proteins. Mean absolute error (MAE) between raw data CVs and average posterior predictive CV are reported. **b**, MAE between held out data and posterior predictive mean separated by genes and proteins for each model and dataset. **c**, Calibration error of held-out data reported separately for genes and proteins. **d**, Stability of estimate for background probability for each cell and protein with respect to default parameters on PBMC10k dataset ($n = 5$ model runs for each of the $n = 14$ proteins). **e**, Held-out marginal log likelihood on the PBMC10k dataset across latent dimensions ($n = 5$ model runs). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range.



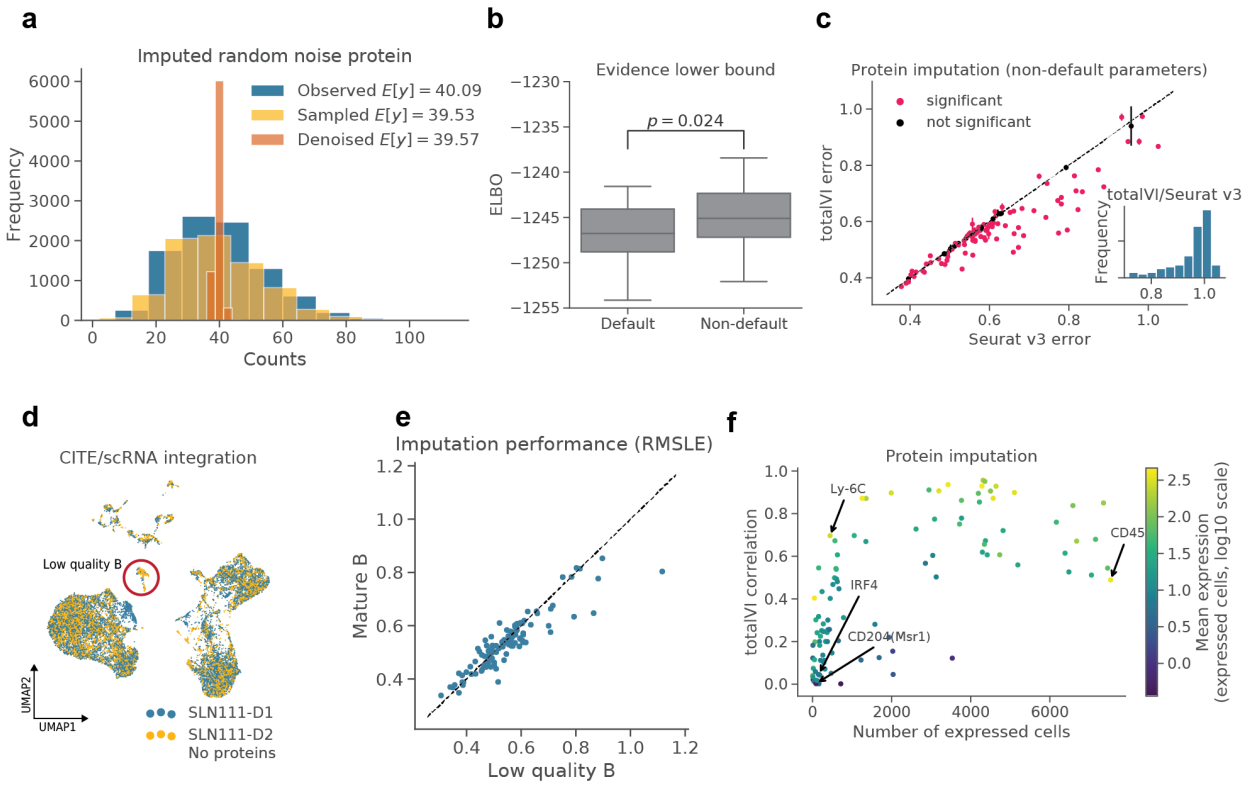
Supplementary Figure 9: Feature autocorrelation across latent space representations. **a, b**, Geary's C calculated in three latent spaces for each feature across all cells in the SLN111-D1 dataset. p-values indicate Wilcoxon rank sum test (two-sided). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **(a)**, Geary's C for each protein ($\log(\text{protein counts} + 1)$, $n = 110$ proteins). In the RNA only vs protein only latent space comparison, $p = 4e - 6$. **(b)**, Geary's C for each gene (log library-size normalized; $n = 4005$ genes). In the RNA only vs protein only comparison, $p = 2e - 8$. For the totalVI joint latent space vs protein only comparison, $p = 1e - 12$. **c, d**, UMAP plots of CD45 isoform expression in the PBMC10k dataset in the scVI **(c)** or totalVI **(d)** latent space. CD45RA and CD45RO proteins are $\log(\text{protein counts} + 1)$ and *PTPRC* RNA is log library-size normalized. **e**, Geary's C calculated for each protein ($\log(\text{protein counts} + 1)$, $n = 14$ proteins) in the totalVI and scVI latent spaces depicted in (c, d). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range.



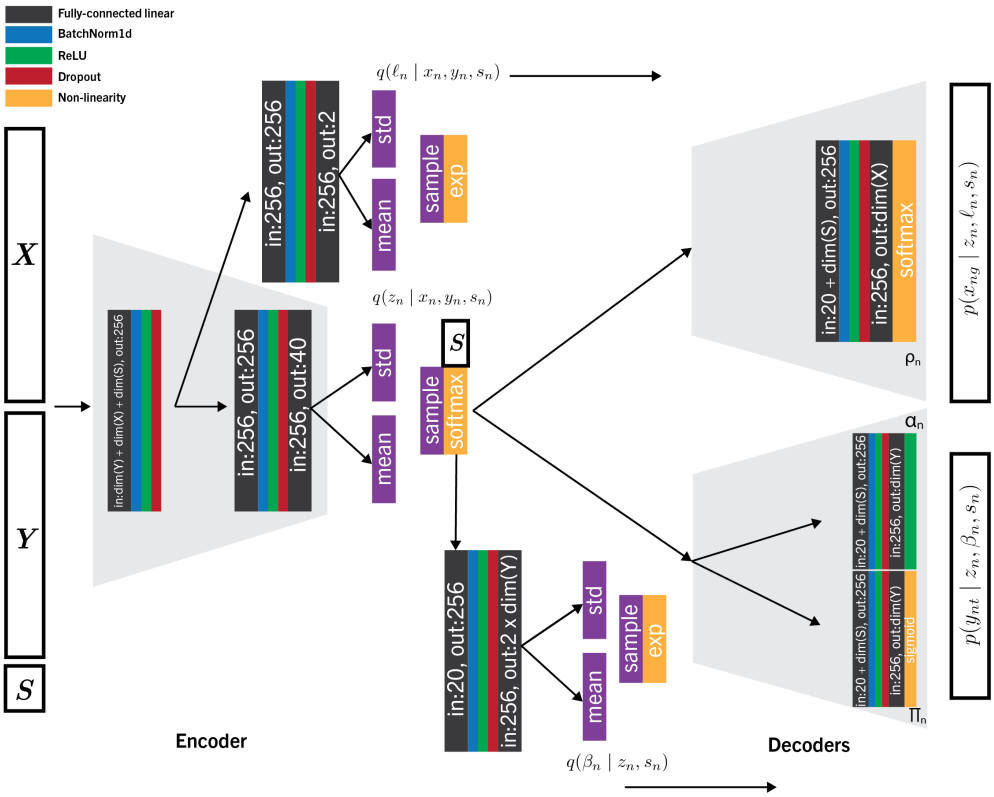
Supplementary Figure 10: Protein and RNA library sizes in the SLN111-D1 dataset. a, b, UMAP plots of SLN111-D1 cells in the SLN-all latent space colored by **(a)** log(protein library size) and **(b)** log(RNA library size). **c**, RNA library size vs protein library size (log scale) for each cell in the SLN111-D1 dataset. **d**, Pearson correlation of RNA and protein library sizes (log scale) by cell type ($n = 26$ cell types) for each cell type depicted in Fig. 4a excluding plasma B (two cells). Box plot indicates the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **e-g**, Protein **(e)** and RNA **(f)** library sizes (log scale) for transitional and mature B cells in the SLN111-D1 dataset. **(g)** RNA library size (log scale) of the genes encoding the top 10 differentially expressed protein markers of mature B cells relative to transitional B cells. **h**, Protein library size in transitional and mature B cells separated by the top 10 differentially expressed markers of mature B cells and the remainder of proteins in the SLN111-D1 dataset. Data are presented as mean values +/- SEM ($n = 867$ transitional B cells and $n = 2,733$ mature B cells). **i**, Geary's C of library size computed per cluster in the SLN-all dataset compared across latent spaces. Left: RNA library size in the RNA-only scVI latent space vs protein library size in the totalVI latent space. Right: RNA library size in the totalVI latent space vs protein library size in the totalVI latent space.



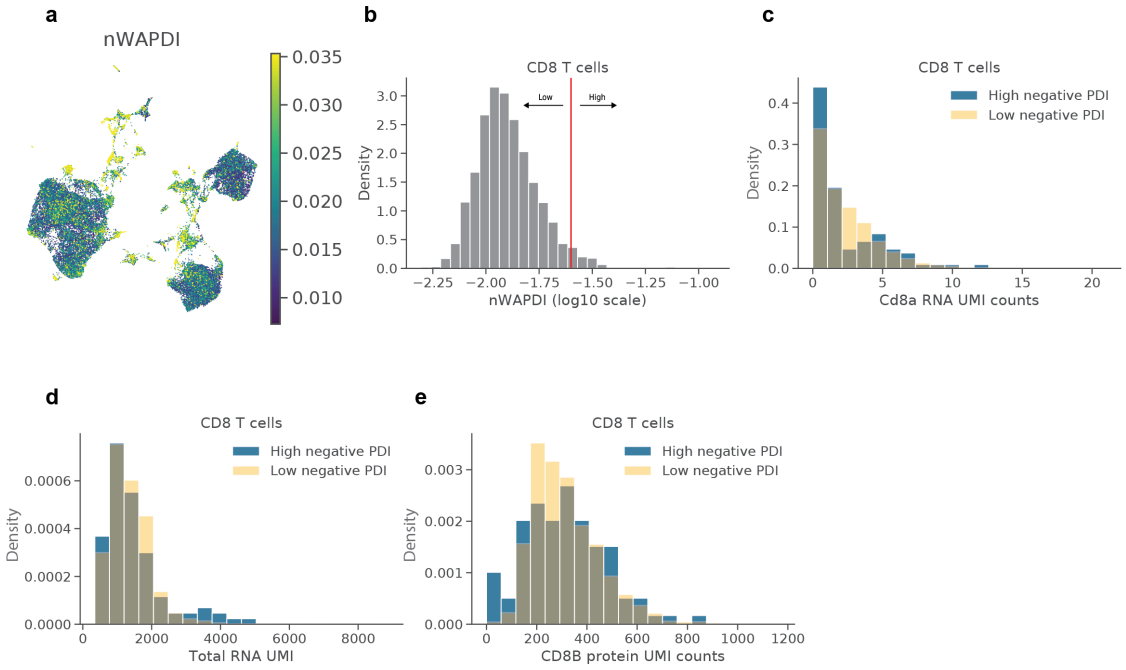
Supplementary Figure 11: Evaluation of totalVI across choice of protein likelihood. a, Posterior predictive check of coefficient of variation (CV) of genes and proteins. For totalVI with Poisson, negative binomial (NB), negative binomial mixture (NB Mixture), and negative binomial mixture with global background and library size correction (Constant BG) protein likelihood, the average coefficient of variation from posterior predictive samples was computed for each feature. Violin plots summarize the distribution of CVs for genes and proteins. Mean absolute error (MAE) between raw data CVs and average posterior predictive CV are reported. **b**, MAE between held out data and posterior predictive mean separated by genes and proteins for each model and dataset. **c**, Calibration error of held-out data reported separately for genes and proteins.



Supplementary Figure 12: Extended analysis of missing protein imputation. a, Distribution of observed protein counts (blue), totalVI imputed protein counts (denoised, orange), and samples from the totalVI imputed distribution (sampled, yellow). The random protein (observed, blue) is a simulated protein added to SLN111-D1 in the imputation task. **b,** Evidence lower bound for SLN imputation task using default parameters and updated (non-default) parameters across $n = 30$ trials. Significance was assessed with a two-sided Welch's t-test. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **c,** totalVI imputation performance versus Seurat v3 imputation performance using non-default totalVI parameters. totalVI performance per protein is presented as mean RMSLE with error bars representing 95% confidence intervals of the mean estimate ($n = 30$ model initializations). Proteins colored in black are not significantly different in performance, while those in red are significantly different (two-sided Student's t-test, BH-adjusted p-value < 0.05). Inset displays ratio in performance between totalVI and Seurat v3. **d,** Reproduction of the UMAP in Figure 3f. **e,** Imputation performance (root mean squared log error) for each protein using only cells annotated as low quality B versus those annotated as mature B in SLN111-D2. **f,** totalVI imputation accuracy (Pearson correlation, log scale) versus number of expressed cells, which was estimated using the totalVI foreground probability ($\pi_{nt} < 0.5$). Points (proteins) are colored by mean expression in expressing cells.



Supplementary Figure 13: totalVI neural network architecture.



Supplementary Figure 14: Evaluation of SLN-all fit with negative widely applicable posterior dispersion indices (nWAPDI). a, nWAPDI computed for each cell in SLN-all on UMAP projection of the totalVI latent space. **b,** Distribution of nWAPDI for CD8 T cells, with decision boundary for cells marked as low nWAPDI and high nWAPDI. **c**, *Cd8a* expression (raw UMI counts) for CD8 T cells with low and high nWAPDI. **d**, RNA library size for same subpopulation split by low/high nWAPDI. **e,** Raw UMI counts of CD8B protein expression using same subpopulation and split.

Supplementary Note 1

On choosing the number of latent dimensions

The latent representation of a cell in totalVI, z_n, is critical to many of the tasks performed in this manuscript. Here we sought to understand the impact of the number of dimensions chosen for z_n on totalVI's modeling capabilities. To do so, we ran totalVI with 5, 10, 20, and 100 latent dimensions, and tested each model in our posterior predictive checking framework. For metrics computed on the training data and held-out data, totalVI's performance was fairly stable across this choice of hyperparameter (Supplementary Figure 8a-c). We note that the denoising component of totalVI also depends on z_n, so we measured the stability of the protein background probability estimate on the PBMC10k dataset across latent dimension choices and for five initializations each, using the default parameters (20 latent dimensions) as a baseline. We again found this estimate to be stable with respect to our default parameters (Supplementary Figure 8d). Finally, we measured the held out marginal log likelihood of the PBMC10k dataset across latent dimension choices. The marginal log likelihood improved for 100 dimensions, which could be due to increased capacity (Supplementary Figure 8e). We note that the relationship between the marginal log likelihood and all other downstream tasks is not well understood. While the marginal log likelihood on held-out data can be used a heuristic for choosing the number of dimensions, it remains unclear in this analysis what impact this will have on interpretation. This phenomenon of stability with respect to the number of latent dimension has also been reported by others [30–32].

Overall, we recommend using the default choice of 20 latent dimensions, which sits between the number used in the scVI [30] publication (10), and the number of principal components used in standard pipelines like Scanpy (about 30), and was used to achieve state-of-the-art results throughout this manuscript. Furthermore, the choice of 20 allowed us to interpret the totalVI latent space with archetypal analysis. If the number of dimensions is much smaller than the number of cell types, archetypal analysis may be more difficult.

Supplementary Note 2

On combining RNA and protein information in a joint latent space

To further explore the latent representation of totalVI's joint model and how it combines RNA and protein information to define cell-cell similarities, we considered the autocorrelation of each feature as measured by Geary's C [33]. We made comparisons to an RNA-only latent representation produced by scVI [30] and to a protein-only latent representation produced by PCA (each with 20 latent dimensions). In the SLN111-D1 dataset, we found that totalVI increased the autocorrelation (lower Geary's C) of protein features relative to RNA only ($p = 0.028$, Wilcoxon rank sum test (two-sided)) (Supplementary Figure 9a). Simultaneously, totalVI increased the autocorrelation of RNA features relative to protein only ($p< 0.001$) (Supplementary Figure 9b). Notably, the autocorrelation of RNA features in the totalVI latent space was not significantly different from that of scVI ($p = 0.102$), implying that the joint latent space of totalVI does not sacrifice the richness of information in the RNA data in order to incorporate protein information.

Combining RNA and protein information into a joint latent representation can be particularly beneficial when the protein data includes isoforms not detectable by RNA sequencing that measures transcript counts rather than full length molecules. For example, the CD45 isoforms CD45RA and CD45RO are commonly used surface markers to distinguish naive from memory human T cells. In an RNA-based analysis, it is possible to define a latent space based on RNA followed by the annotation of cells with their protein isoform expression (Supplementary Figure 9c). However, a joint analysis can make use of protein isoform expression along with transcriptome measurements to define a latent space and cell-cell similarities (Supplementary Figure 9d). As expected, totalVI's joint model increases the autocorrelation of proteins relative to an RNA-only analysis (Supplementary Figure 9e), indicating that totalVI incorporates isoform information into the latent space.

Supplementary Note 3

Protein considerations

Experimental considerations Sources of technical variation in CITE-seq experiments, particularly protein background, are dependent on the experimental method itself. There are a number of potential experimental sources of background. We primarily discussed ambient antibodies and non-specific antibody binding. Another potential source of background could arise from oligonucleotide barcodes that become dissociated from their conjugated antibody. Similarly to barcoded antibodies, ambient oligonucleotide barcodes could contaminate cell-containing droplets or could non-specifically bind to the cell surface. In this study, we do not distinguish between background due to antibodies or background due to free oligonucleotide barcodes.

Although in our experiments we used the standard CITE-seq protocol, there are a number of protocol modifications that could change the amount of background. For instance, increasing the number of washes after staining cells with antibodies could reduce ambient background. Alternatively, a buffer modification could reduce the amount of non-specific binding. Both washing and blocking are frequently considered in flow cytometry protocol designs. However, implementing these protocol changes in an effort to eliminate background could come with trade-offs; reducing background by washing and blocking would likely reduce true signal by reducing total cell numbers or blocking specific binding, respectively.

Another common experimental practice to modulate the amount of background is antibody titration, meaning that different antibodies are added to the experiment at different concentrations. At the optimal titration, an antibody would have the maximal signal-to-noise ratio. This would require the antibody to be present at a sufficient concentration to specifically bind its target protein and generate a detectable signal, but not at so high a concentration as to increase protein background by binding non-specifically or remaining at high concentration in the ambient solution. In a CITE-seq experiment, it is possible that the recommended antibody concentration is too low to detect foreground signal from a given protein. Finding the optimal concentration for each antibody might be challenging, since the optimal concentration might be different for different cell types or experimental systems. If titrations are modified per antibody, there are a few points to consider. When antibodies are titrated at different concentrations, it becomes infeasible to quantify absolute protein levels. For example, it would not be possible to determine if one protein was expressed at a higher level than another protein. In addition, even if every protein in the cell were measured with a theoretical unbiased antibody panel, the sum of all protein counts from a cell (referred to as the protein library size) could not serve as a meaningful estimate of total protein molecules in a cell or cell size because the relative amounts of each protein have been manipulated.

For proteins that are expressed at low levels or are only expressed in rare cell types, it might be necessary to increase sequencing depth to increase the sensitivity to detect these molecules. In addition, sequencing depth could play a role in the ability of totalVI to decouple protein foreground and background (i.e., with more counts, it might become easier to separate what appear to be overlapping foreground and background distributions). Determining the optimal sequencing depth for

protein panels could be an important cost consideration in CITE-seq experiments, particularly as the size of protein panels increases. Since in the CITE-seq protocol RNA and protein libraries are prepared independently, future work could determine the value of these two molecules in various downstream analysis tasks to make recommendations for sequencing depth for each library.

Because the barcoded antibodies used in this study came from clones that have been previously validated, we were surprised to find that some common protein markers (e.g., IgM, CCR7) appeared to have little or no signal. Aside from the consideration of titration and sequencing depth discussed above, an additional explanation could be the uniform staining conditions for all antibodies in the CITE-seq panel simultaneously. For example, the chemokine receptor CCR7 is a well-documented marker in T cells and typically requires staining at higher temperature and for longer times than other antibodies due to its constant cycling onto and off of the cell surface. For future CITE-seq experiments, it might be worthwhile to consider the optimal staining conditions (e.g., time and temperature) for each antibody independently rather than staining with all antibodies at once.

Modeling considerations Guided by the points raised above, we considered a variety of protein likelihoods before settling on the version used in this manuscript. Among our considerations were the interpretability of the parameters as well as how well the likelihood captured our view on the CITE-seq protein data generating process.

In our modeling and analysis, we considered whether protein library size should be taken into account. For example, we considered models that included a latent variable for the protein library size (analogous to ℓ_n for RNA). Here, we discuss why protein library size does not convey the same information as RNA library size and is thus not treated the same in the totalVI model.

In scRNA-seq data, we consider library size to be a nuisance factor that is reflective of a combination of sequencing depth and cell size. Although the number of RNA transcripts and the number protein molecules both scale with the size of a cell [34], the unbiased sampling of the transcriptome is much more likely to approximate the relative size of a cell than the sampling of a limited selection of proteins on the cell surface. In CITE-seq experiments where only selected markers are measured, there is no guarantee that the markers selected are representative of the total protein content in the cell. For example, a cell with few detected protein counts might in reality express other unmeasured proteins at higher levels, meaning this cell's total counts reflect the selection of markers rather than reflecting nuisance variation like sequencing depth or cell size. Therefore, treating protein library size as a nuisance factor does not make sense in this context. Because measured proteins are biased in this manner, we do not assume that the protein data is compositional, as is assumed by other methods that use a centered log ratio (CLR) transformation for normalization per cell [35].

To further explore the effects of library size, we considered protein and RNA library sizes in the SLN111-D1 dataset (Supplementary Fig. 10a-c). While RNA and protein library sizes (log scale) were positively correlated (Pearson correlation of 0.54), this correlation varied widely by cell type (Supplementary Fig. 10d). This observation suggested that within some cell types, protein library size might approximate a cell's relative size similarly to RNA library size. However, in other cell

types, bias in the antibody panel results in total protein expression that reflects biological differences in the measured proteins rather than a global measurement of protein content. The difference between RNA and protein library size can be observed when comparing transitional and mature B cells, which have significantly different protein library sizes (Welch's t-test, $p < 0.01$), but not significantly different RNA library sizes (Welch's t-test, $p = 0.098$) (Supplementary Fig. 10e-f). The difference in protein library size between these cell types can largely be explained by differences in expression of measured mature B cell markers: 74% of the mean difference in protein library size is driven by the 10 most differentially expressed proteins in mature vs transitional B cells (Methods), including known markers like IgD and CD23 (Supplementary Fig. 10g, h). Therefore, normalizing each cell by its protein library size would remove biological differences and introduce additional bias based on the selection of measured proteins. Finally, we considered how library size is reflected in the latent space by calculating the autocorrelation of library size per cluster as measured by Geary's C [33]. The autocorrelation of protein library size in the totalVI latent space was similar to the autocorrelation of RNA library size in either the RNA-only latent space computed by scVI or in the totalVI latent space, both of which remove the effect of RNA library size (Supplementary Fig. 10i). This finding further supported the notion that the totalVI latent space was not unduly biased by protein library size. In the future, more unbiased protein panels might necessitate further consideration of protein library size in CITE-seq data analysis.

We also considered whether RNA background should be addressed similarly to protein background in our model. In addition to our observations that levels of background RNA were far lower than protein background (Extended Data Fig. 3d-f), we considered estimates of UMI counts due to background for RNA and protein. For RNA background, we refer to the DecontX method that estimates and removes contamination of ambient RNA in scRNA-seq data [36]. According to the DecontX study, in three scRNA-seq datasets of different tissues collected using the 10x v3 platform, the median percentage of RNA UMI counts estimated to be derived from ambient RNA was 0.03% (brain), 0.12% (heart), and 0.56% (PBMC) [36]. For comparison, we estimated the number of protein UMI counts that could be attributed to background in each cell by summing all protein UMI counts with a foreground probability < 0.5 and dividing by the total protein UMI counts. For the data reported in this study (also collected using the 10x v3 platform), the estimated median percentage of protein background UMI counts across all cells in SLN-all was 12.71%. By these estimates, RNA background UMI counts are low (<1%, approximately two orders of magnitude lower than protein background UMI counts). We therefore chose not to explicitly model RNA background in totalVI.

Additionally, we considered alternative models to decouple protein background. Initially, we attempted to use simple likelihoods like Poisson or negative binomial, or models that assume every cell receives the same distribution of protein background scaled by some cell-specific scalar. However, we found these models inadequate for decoupling the protein signal, which again suggests that ambient antibodies can not fully explain the protein background, and that our proposed negative binomial mixture fits the data better (Supplementary Fig. 11).

The likelihood we used in this manuscript,

$$y_{nt} \mid z_n, \beta_n, s_n \sim \text{NegativeBinomialMixture}(\beta_{nt}, \beta_{nt}\alpha_{nt}, \pi_{nt}), \tag{1}$$

also has some important downstream considerations. First, the mixture assumes that the observed counts for a given cell n and protein t are generated from either the background component (with probability π_{nt}) or the foreground component (with probability $1 - \pi_{nt}$). Despite the fact that the background mean parameter β_{nt} appears in the foreground mean $\beta_{nt}\alpha_{nt}$, this likelihood does not allow us to correct the foreground for possible background contamination. Here, the double usage of β_{nt} is to help identify the mixture model. In other words, we cannot "subtract the background" from y_{nt} that are determined to be in the foreground. Perhaps this limitation could be addressed in future work, in which different latent variables are associated with local and global sources of background; though this will require greater understanding of the experimental mechanisms previously discussed.

As a final point, the negative binomial mixture we used in this manuscript may be less suitable in the case where the dataset contains a homogeneous population of cells. This is due to there being a lack of cells that would be considered background for the proteins. In this case, one can either use the mean of the negative binomial mixture (without expected background subtraction) for downstream analysis, or optionally use the negative binomial distribution as the protein likelihood (an option in totalVI).

Supplementary Note 4

On imputing missing proteins

In our analysis, we have shown that totalVI can accurately predict missing proteins, with improvement over the predictions of Seurat v3. Here we further discuss the merits and limitations of missing protein imputation.

We first sought to quantify any bias in totalVI's protein imputation. As an illustrative example, we simulated a protein as independently and identically distributed from a negative binomial distribution
(`np.random.negative_binomial(10, 0.2)`.). We concatenated this protein to the others and reran the imputation example used in Fig. 3 on our spleen and lymph node data (SLN111-D1 and SLN111-D2 with no proteins). The default totalVI prediction is displayed as the orange histogram in Figure 12a. As the totalVI denoised counts represent the expected value, we should expect that if totalVI does not produce biased predictions that the values pileup around 40, the expected value of the simulated negative binomial data (in blue). Because totalVI fits the full distribution, instead of reporting the expectation, we can sample from it. From this sampling, we observed that the samples closely matched the empirical distribution of the random protein. Taken together, in this particular experiment, totalVI's predictions have little bias, but the degree of noise in the proteins sets a ceiling on the imputation performance of any algorithm.

We also sought to understand the impact of totalVI's hyperparmeters on the imputation task. Throughout the manuscript we used a set of default hyperparameters, so as the provide reasonable defaults that would yield good performance across many tasks and datasets with distinct characteristics. However, the task of protein imputation requires the model to not backpropagate certain nodes in the network as well as to handle zeros for missing data; so, it is reasonable that another set of hyperparameters could yield improved performance on this task. Guided by maximizing the evidence lower bound, we found another set of hyperparameters that differed from the defaults only in the number of hidden layers (two, relative the default one), and a reduction of the learning rate from `4e-3` to `2e-3`. We again performed the imputation task on the spleen and lymph node data over 30 initializations. We found the evidence lower bound of the data to be significantly higher with the non-default parameters (Welch's t-test, p-value < 0.05; Supplementary Figure 12b). Furthermore, we found 89 proteins to be significantly different in their root mean squared logarithmic error to that of Seurat v3, 68 of which had better performance for totalVI (Supplementary Figure 12c). This is an improvement over the default parameters, though it is quite minor, indicating that the task is somewhat robust to these choices.

Finally, we consider the usage of totalVI imputed proteins as proxies for real observed protein measurements. In our analysis, totalVI had good imputation accuracy for many proteins. One relevant consideration is the quality of the proteins in the reference dataset. Imputation performance may suffer due to poor antibodies or lack of biological expression, and it will not always be straightforward to understand which effect is being observed, especially as CITE-seq panels grow to be more unbiased. Another consideration is the extent to which the reference and query datasets

share underlying biology. totalVI relies on the assumption that a cell with no observed proteins will have cells from the dataset with observed protein expression in its neighborhood in the latent space. If a cell type is missing from the reference dataset that is observed in the query dataset, we do not expect good imputation performance; though it may depend how biologically similar the missing cell type is to the others in the reference dataset.

For example, the UMAP plot in Fig. 3f (reproduced here as Supplementary Figure 12d) revealed a small subpopulation of cells from SLN111-D2 that did not integrate well with the cells of SLN111-D1. Based on our annotations, these cells were overwhelmingly from a subpopulation of low quality B cells (high mitochondrial content) that were enriched in the SLN111-D2 dataset relative to the remaining spleen and lymph node datasets, explaining the lack of mixing. We compared the imputation performance via correlation for these cells relative to a related population of mature B cells, and found the performance tended to be similar in the low quality B cells (Supplementary Figure 12e). However, this result may be due to the relative similarity of these low quality B cells to the higher quality mature B cells, coupled with the fact that the RNA and not the protein data were low quality. Generally, we believe that imputation performance will be better for those cells that mix well with CITE-seq cells. The potential pitfall of poor imputation quality can be mitigated by empirically quantifying the presence of CITE-seq cells in a cell's neighborhood. While imputed proteins can be used in downstream analysis to generate hypotheses, these predictions should be validated experimentally before drawing concrete biological conclusions.

We also note that totalVI is capable of imputing the expression of proteins that are expressed in rare cell types (Supplementary Figure 12f). Here we investigated the Pearson correlation as an imputation performance metric, which allowed us to evaluate proteins on the same scale, as the RMSLE depends on the range of expression values and mean of the protein. In Supplementary Figure 12f, we can see that totalVI imputes Ly-6C well, which is expressed in smaller subpopulations of CD8 T cells and monocytes. In contrast, IRF4 is a negative control protein (intracellular) that was detected in very few cells (likely technical artifacts), explaining its low correlation.

Supplementary Note 5

Integrating out latent variables

Here we show that if

$$w \sim \text{Gamma}(\theta, \ell\rho) \tag{2}$$
$$x \mid w \sim \text{Poisson}(w) \tag{3}$$

then $x \sim \text{NegativeBinomial}(\ell\rho, \theta)$. Note that we have dropped all subscripts and each variable here is a scalar. While we parameterize the Gamma with its shape and mean, a more conventional form is with its shape and rate, so $w \sim \text{Gamma}(\theta, \theta/(\ell\rho))$

$$p(x) = \int_0^\infty p(x \mid w)p(w)dw \tag{4}$$
$$= \int_0^\infty \frac{w^x e^{-w}}{\Gamma(x+1)} \frac{(\theta/(\ell\rho))^\theta}{\Gamma(\theta)} w^{\theta-1} e^{-\theta w/(\ell\rho)} dw \tag{5}$$
$$= \frac{(\theta/(\ell\rho))^\theta}{\Gamma(x+1)\Gamma(\theta)} \int_0^\infty w^{x+\theta-1} e^{-(1+\theta/(\ell\rho))w} dw \tag{6}$$
$$= \frac{(\theta/(\ell\rho))^\theta}{\Gamma(x+1)\Gamma(\theta)} \frac{\Gamma(x+\theta)}{(1+\theta/(\ell\rho))^{x+\theta}} \tag{7}$$
$$= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)} \left(\frac{\theta/(\ell\rho)}{1+\theta/(\ell\rho)}\right)^\theta \left(\frac{1}{1+\theta/(\ell\rho)}\right)^x \tag{8}$$
$$= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)} \left(\frac{\theta}{\ell\rho+\theta}\right)^\theta \left(\frac{\ell\rho}{\ell\rho+\theta}\right)^x \tag{9}$$

In the fourth line, we use the fact that the integrand is an unnormalized gamma distribution. The final line is a negative binomial distribution with mean $\ell\rho$ and inverse dispersion θ. Therefore, we have a direct link between the parameters of the negative binomial and the underlying parameters of the Poisson rate. Finally, we note that we could have written $w \sim \text{Gamma}(\theta, \rho)$ and $x \mid w \sim \text{Poisson}(\ell w)$ and achieved the same result.

Supplementary Note 6

totalVI implementation details

Evidence lower bound derivation Here we derive the Evidence Lower Bound (ELBO), which is ultimately used in optimizing the model and variational parameters. For shorthand, we drop subscripts and inference and generative parameters ν and η. The joint likelihood based on the totalVI generative model for a single cell factorizes as

$$p(x, y, \beta, z, \ell \mid s) = p(x \mid z, \ell, s)p(y \mid \beta, z, s)p(\beta \mid s)p(z)p(\ell \mid s). \tag{10}$$

In the model specification, we use the latent variable $z \sim \text{LogisticNormal}(0, I)$. Here we use the logistic normal definition of [37, 38], in which a normal random variable $\delta \sim \text{Normal}(0, I)$ is transformed by a softmax function, embedding the random variable in the simplex. Thus, $z = \text{softmax}(\delta)$. However, the softmax function is not invertible, so for simplicity we consider the underlying latent variable δ. In this setting, z, which is ultimately the input to the decoder, is treated as a likelihood parameter. Therefore, we can rewrite the joint likelihood as

$$p(x, y, \beta, \delta, \ell \mid s) = p(x \mid \delta, \ell, s)p(y \mid \beta, \delta, s)p(\beta \mid s)p(\delta)p(\ell \mid s). \tag{11}$$

To perform variational inference, we define the variational posterior distribution as

$$q(\beta, \delta, \ell \mid x, y, s) = q(\beta \mid \delta, s)q(\delta \mid x, y, s)q(\ell \mid x, y, s). \tag{12}$$

The ELBO is derived using Jensen's inequality. We use the shorthand notation $q(\beta, \delta, \ell) = q(\beta, \delta, \ell \mid x, y, s)$.

$$\begin{aligned}
\log p(x, y \mid s) =& \log \mathbb{E}_{q(\beta,\delta,\ell)} \left[\frac{p(x, y, \beta, \delta, \ell \mid s)}{q(\beta, \delta, \ell)} \right] && (13)\\
\geq& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log \frac{p(x, y, \beta, \delta, \ell \mid s)}{q(\beta, \delta, \ell)} \right] && (14)\\
=& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log p(x, y \mid \beta, \delta, \ell, s) \right] + \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log \frac{p(\beta \mid s)p(\delta)p(\ell \mid s)}{q(\beta, \delta, \ell)} \right] && (15)\\
=& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log p(x, y \mid \beta, \delta, \ell, s) \right] - \text{KL}\left(q(\ell) \parallel p(\ell \mid s)\right) - \text{KL}\left(q(\delta) \parallel p(\delta)\right) && (16)\\
& - \mathbb{E}_{q(\delta)} \left[\text{KL}\left(q(\beta) \parallel p(\beta \mid s)\right) \right] && (17)
\end{aligned}$$

To compute the KL divergences of lognormal random variables we note that the KL divergence is invariant to invertible transformations, so the KL can be computed in closed form using the KL between normal random variables. The log likelihood of a negative binomial mixture distribution is computed using numerically stable functions in Pytorch (see below). The ELBO derived here is amenable to the reparameterization trick used to train VAEs [39]. Estimates of the expectations in the ELBO are taken via Monte Carlo and noisy gradients of the ELBO are used in a stochastic optimization scheme. A sketch of the inference procedure for totalVI is in Algorithm 2.

Approximate posterior specification The approximate posterior distributions are specified by neural networks. In particular, one neural network takes as input the triple (x, y, s) and outputs the parameters of $q(\delta \mid x, y, s)q(\ell \mid x, y, s)$. An additional neural network maps (δ, s) to the mean and variance parameters of $q(\beta \mid z, s)$ through z. The variational distributions match their priors in family (e.g., $q(\delta \mid x, y, s)$ is a Gaussian with diagonal covariance matrix). A posterior draw of z, which we used as input to clustering and visualization algorithms, as well as used for archetypal analysis is then obtained by

1. Draw δ from $q(\delta \mid x, y, s)$
2. Set $z = \text{softmax}(\delta)$

The mean of $q(z \mid x, y, s)$ can be computed using Monte Carlo integration.

Neural networks The encoder neural network has one shared hidden layer of 256 nodes followed by a layer of 512 nodes. The output of the 512 nodes are split in half and are used as input for linear layers that parameterize $q(\delta \mid x, y, s)$ and $q(\ell \mid x, y, s)$, respectively. The final encoder neural network of one hidden layer and 256 hidden nodes takes as input (z, s) and outputs the parameters of $q(\beta \mid \delta, s)$. The parameters of δ are 20-dimensional mean and variance parameters. The decoder consists of three individual neural networks each with one hidden layer and 256 nodes. The first maps to the parameters of the mean of the RNA likelihood (ρ_n). The second maps to the foreground mean of the protein likelihood (α_n). The third maps to the mixing parameter of the protein likelihood mixture (π_n). Each of these decoder networks takes as input (z, s). Furthermore, (z, s) are reinjected at each hidden layer. All neural networks use batch normalization [40], dropout regularization [41], and ReLU activations in hidden layers. The model parameters θ, ϕ, c, and d are treated as global neural network parameters, optimized to maximize the ELBO. A schematic of the architecture is in Supplementary Figure 13. We note that the architecture (fully-connected, one to two hidden layers, ReLU activation, etc.) is similar to other VAE/autoencoder models used for single-cell data including the scVI and DCA models [30, 31].

Hyperparameters The neural network architecture previously described was used throughout this manuscript without modification. There are a number of other hyperparameters used to train neural networks that we also held constant in all experiments. This includes the learning rate of the optimizer (`lr=4e-3`), the size of the training test (90%), the KL warmup scheme ($0.75\times$`NumCells` minibatches, or 213 epochs with the fixed training set size of 90%), and the number of training epochs (500 epochs). An early stopping scheme is performed with respect to the 10% of cells in the test set. If there is no improvement of the held-out ELBO of the test set with 30 epochs, the learning rate is multiplied by 0.6. If there is no improvement after 45 epochs, the inference procedure is stopped.

Numerical considerations for the negative binomial mixture distribution For a single negative binomial mixture component, we use numerically stable functions provided by PyTorch (e.g.,

Algorithm 2: Inference for totalVI

Initialize inverse dispersion parameters θ, ϕ, background parameters c_t, d_t and encoder/decoder neural network parameters.
for *iteration* $i = 1, 2, ...,$ **do**
  Randomly choose M cells for mini-batch $\mathcal{C}$
  for *each cell* n *in* $\mathcal{C}$ **do**
    Encode x_n, y_n, s_n to obtain approximate posterior parameters
    Sample z_n, ℓ_n, β_n from approximate posterior $q(\beta_n, \delta_n, \ell_n \mid x_n, y_n, s_n)$
    Decode z_n, s_n to obtain likelihood parameters α_n, π_n, ρ_n
    for *each gene* g **do**
      Compute $\log p(x_{ng} \mid \ell_n, z_n, s_n)$
    for *each protein* t **do**
      Compute $\log p(y_{nt} \mid \beta_{nt}, z_n, s_n)$
  Update parameters using gradient of ELBO estimate

the log gamma function). For a mixture of negative binomials, we rewrite the distribution to use numerically stable functions like `logsumexp` and `softplus`.

Let $p^b(y) = p(y \mid z, \beta, s, v = 1)$ be the probability mass function for the background and $p^f(y) = p(y \mid z, \beta, s, v = 0)$ be the probability mass function for the foreground. Then by integrating over v (recalling that $v \sim \text{Bernoulli}(\pi)$),

$$\log p(y \mid z, \beta, s) = \log \big(\pi p(y \mid z, \beta, s, v = 1) + (1 - \pi) p(y \mid z, \beta, s, v = 0)\big). \tag{18}$$

We now rewrite this in a form more amenable for optimization. We recall from Algorithm 1 that $\pi = h_\pi(z, s; \Omega)$. Thus, with $c(z, s) = \text{logit}(h_\pi(z, s))$, then $\pi = 1/(1 + \exp(-c(z, s)))$. Also, let $\mathcal{S}$ be the softplus function: $x \to \log(1 + e^x)$.

$$\begin{aligned}
\log p(y \mid z, \beta, s) &= \log \big(\pi p^b(y) + (1 - \pi) p^f(y)\big) && (19)\\
&= \log \big(p^f(y) + \pi p^b(y) - \pi p^f(y)\big) && (20)\\
&= \log \left(\frac{p^f(y) + p^f(y) e^{-c(z,s)}}{1 + e^{-c(z,s)}} + \frac{p^b(y) - p^f(y)}{1 + e^{-c(z,s)}}\right) && (21)\\
&= \log \big(p^b(y) + p^f(y) e^{-c(z,s)}\big) - \log \big(1 + e^{-c(z,s)}\big) && (22)\\
&= \log \left(e^{\log p^b(y)} + e^{\log p^f(y)} e^{-c(z,s)}\right) - \log \big(1 + e^{-c(z,s)}\big) && (23)\\
&= \text{logsumexp} \big(\log p^b(y), \log p^f(y) - c(z, s)\big) - \mathcal{S}(-c(z, s)) && (24)
\end{aligned}$$

Supplementary Note 7

Posterior dispersion indices highlight model misfit

Here we describe how we used posterior dispersion indices to further evaluate the totalVI model fit. In this analysis, we used the negative widely applicable dispersion index (nWAPDI) [42]. With this metric, each cell gets a value describing the uncertainty of the likelihood of a particular cell with respect to the latent variables. Higher values indicate that the model is failing to explain the observed expression in a particular cell. The nWAPDI is computed for cell n as

$$\text{nWAPDI(n)} := -\frac{\sigma^2_{\log}(n)}{\log \mu(n)}, \tag{25}$$

where the quantities

$$\mu(n) = \mathbb{E}_{q(z_n,\beta_n,\ell_n)}[p(x_n, y_n \mid z_n, \beta_n, \ell_n, s_n)], \tag{26}$$

$$\sigma^2_{\log}(n) = \mathbb{V}_{q(z_n,\beta_n,\ell_n)}[\log p(x_n, y_n \mid z_n, \beta_n, \ell_n, s_n)], \tag{27}$$

can be computed using samples from the approximate posterior (1000 samples for each cell). We computed the nWAPDI for each cell in the SLN-all model fit (Supplementary Figure 14a). Next, we looked at the distribution of nWAPDI scores inside a homogeneous cell type like CD8 T cells and explored phenotypic differences between outliers and the remaining CD8 T cells. We found that those cells with high nWAPDI tended to have a surprising zero UMI count frequency for key cell type markers like the Cd8a gene and the CD8B protein, and it could not be explained by a difference in library size (Supplementary Figure 14b-e). This indicates that totalVI may not be explaining particularly surprising zeros well and could suggest the use of zero-inflated distributions, though this remains future work.

References

1. Li, B. *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nature Methods* **17,** 793–798 (2020).
2. Kirchner, J. & Bevan, M. J. ITM2A is induced during thymocyte selection and T cell activation and causes downregulation of CD8 when overexpressed in CD4+CD8+ double positive thymocytes. *Journal of Experimental Medicine* **190,** 217–228 (1999).
3. Kreslavsky, T. *et al.* Essential role for the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nature Immunology* **18,** 442–455 (2017).
4. Macias-Garcia, A. *et al.* Ikaros and B1 cells Ikaros is a negative regulator of B1 cell development and function. *Journal of Biological Chemistry* (2016).
5. Shi, Z. *et al.* Human CD8+ CXCR3+ T cells have the same function as murine CD8+ CD122+ Treg. *European Journal of Immunology* **39,** 2106–2119 (2009).
6. Liu, J., Chen, D., Nie, G. D. & Dai, Z. CD8+CD122+ T-cells: A newly emerging regulator with central memory cell phenotypes. *Frontiers in Immunology* **6** (2015).
7. Lai, L., Alaverdi, N., Maltais, L. & Morse, H. C. Immunophenotyping mouse cell surface antigens: Nomenclature and immunophenotyping. *The Journal of Immunology* (1998).
8. Merad, M., Sathe, P., Helft, J., Miller, J. & Mortha, A. The Dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annual Review of Immunology* **31,** 563–604 (2013).
9. Durai, V. & Murphy, K. M. Functions of murine dendritic cells. *Immunity* **45,** 719–736 (2016).
10. Scott, R. E., Ghule, P. N., Stein, J. L. & Stein, G. S. Cell cycle gene expression networks discovered using systems biology: Significance in carcinogenesis. *Journal of Cellular Physiology* **230,** 2533–2542 (2015).
11. Doty, R. T. *et al.* Coordinate expression of heme and globin is essential for effective erythropoiesis. *Journal of Clinical Investigation* **125,** 4681–4691 (2015).
12. Sagar *et al.* Deciphering the regulatory landscape of $\gamma\delta$ T Cell development by single-cell RNA-sequencing. *bioRxiv*. `https://doi.org/10.1101/478529` (2018).
13. Burmeister, Y. *et al.* ICOS controls the pool size of effector-memory and regulatory T cells. *The Journal of Immunology* **180,** 774–782 (2008).
14. Chen, X. & Oppenheim, J. J. Resolving the identity myth: key markers of functional CD4 + FoxP3 + regulatory T cells. *International Immunopharmacology* (2011).
15. Sunderkötter, C. *et al.* Subpopulations of mouse blood monocytes differ in maturation stage and inflammatory response. *The Journal of Immunology* **172,** 4410–4417 (2004).
16. Marcovecchio, P. M. *et al.* Scavenger receptor CD36 directs nonclassical monocyte patrolling along the endothelium during early atherogenesis. *Arteriosclerosis, thrombosis, and vascular biology* **37,** 2043–2052 (2017).

17. Loder, F. *et al.* B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. *Journal of Experimental Medicine* **190,** 75–89 (1999).

18. Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage. *Nature Immunology* (2012).

19. Borges Da Silva, H. *et al.* Splenic macrophage subsets and their function during blood-borne infections. *Frontiers in Immunology* **6,** 480 (2015).

20. Tardif, M. R. *et al.* Secretion of S100A8, S100A9, and S100A12 by neutrophils involves reactive oxygen species and potassium efflux. *Journal of Immunology Research* (2015).

21. Fehniger, T. A. *et al.* Acquisition of Murine NK cell cytotoxicity requires the translation of a pre-existing pool of Granzyme B and Perforin mRNAs. *Immunity* **26,** 798–811 (2007).

22. Bendelac, A., Rivera, M. N., Park, S.-H. & Roark, J. H. Mouse CD1-Specific NK1 T Cells: Development, Specificity, and Function. *Annual Review of Immunology* **15,** 535–562 (1997).

23. Zhang, J. *et al.* Characterization of Siglec-H as a novel endocytic receptor expressed on murine plasmacytoid dendritic cell precursors. *Blood* **107,** 3600–3608 (2006).

24. Sawai, C. M. *et al.* Transcription factor Runx2 controls the development and migration of plasmacytoid dendritic cells. *Journal of Experimental Medicine* **210,** 2151–2159 (2013).

25. Castro, C. D. & Flajnik, M. F. Putting J Chain back on the map: How might its expression define plasma cell development? *The Journal of Immunology* **193,** 3248–3255 (2014).

26. Dutta, P. *et al.* Macrophages retain hematopoietic stem cells in the spleen via VCAM-1. *Journal of Experimental Medicine* **212,** 497–512 (2015).

27. Hobeika, E., Dautzenberg, M., Levit-Zerdoun, E., Pelanda, R. & Reth, M. Conditional selection of B cells in mice with an inducible B cell development. *Frontiers in Immunology* **9,** 1806 (2018).

28. Thomas, M. D., Srivastava, B. & Allman, D. Regulation of peripheral B cell maturation. *Cellular Immunology* **239,** 92–102 (2006).

29. Roncarolo, M.-G. & Gregori, S. Is FOXP3 a bona fide marker for human regulatory T cells? *European Journal of Immunology* **38,** 925–927 (2008).

30. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15,** 1053–1058 (2018).

31. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10,** 390 (2019).

32. Xu, C. *et al.* Harmonization and annotation of single-cell transcriptomics data with deep generative models. *bioRxiv,* 532895 (2019).

33. DeTomaso, D. *et al.* Functional interpretation of single cell similarity maps. *Nature Communications* (2019).

34. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends in Genetics* **28,** 560–565 (2012).

35. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* (2017).

36. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology* **21,** 57 (2020).

37. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *The Annals of Applied Statistics* **1,** 17–35 (2007).

38. Dieng, A. B., Ruiz, F. J. R. & Blei, D. M. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics* **8,** 439–453 (2020).

39. Kingma, D. P. & Welling, M. *Auto-Encoding variational Bayes* in *International Conference on Learning Representations* (2014).

40. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* (2015).

41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15,** 1929–1958 (2014).

42. Kucukelbir, A., Wang, Y. & Blei, D. M. *Evaluating Bayesian models with posterior dispersion indices* in *International Conference on Machine Learning* (2017).