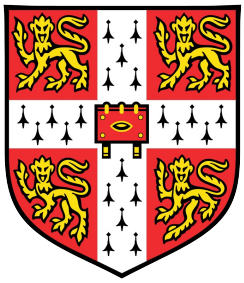
Systems immunology frameworks link multicellular immune perturbation phenotypes and setpoints to response outcomes



Matthew P. Mulè

Cambridge Institute of Therapeutic Immunology & Infectious Disease

Department of Medicine

University of Cambridge

Supervisors: Professors Kenneth G.C. Smith & John S. Tsang

This dissertation is submitted for the degree of

Doctor of Philosophy

Sidney Sussex College September 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 65,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Matthew Mulè September 2022

Systems immunology frameworks link multicellular immune perturbation phenotypes and setpoints to response outcomes

Matthew P. Mulè

Abstract

This thesis develops frameworks for using top-down systems biology approaches with multiomic single cell technology to understand variation in human immune system response outcomes. We integrate human population, cell subset and single cell variations in molecular phenotypes which give rise to baseline setpoints, are perturbed by vaccination or drug treatment, and are linked to emergent response / clinical outcomes. In Chapter 2, we dissect noise sources in data derived from new methods which simultaneously measure protein and mRNA in single cells (e.g., CITE-seq). After identifying two main sources of noise, we develop an open source software method for normalizing and denoising CITE-seq protein data. We then develop a computational framework for analyzing the effects of timed immune system perturbations applied to human cohorts profiled with multimodal single cell technology. Chapter 3 applies these methods on a human vaccination cohort profiled using CITE-seq. We first define highly interpretable immune cell subsets using unsupervised clustering based on the denoised protein data, then contrast the transcriptome pathways within these subsets that are differentially induced by vaccines formulated with and without an adjuvant. These robust phenotypes are further interpreted using single cell computational reconstructions of cell states. Using these comparative analyses, along with unbiased analysis of baseline phenotypes linked to antibody response, we identify a multicellular "naturally adjuvanted" human immune system setpoint more poised to respond to vaccination. Chapter 4 applies the methods developed above to a cohort of cancer patients treated with immune checkpoint inhibitors. In this work, we identify multicellular baseline setpoints linked to development of immune related adverse events after treatment which are uncoupled from the phenotypes induced by treatment. Together, these approaches help advance a quantitative, predictive understanding of human immune system variation, and pave the way for further human perturbation cohort studies across biological disciplines.

Acknowledgements

I first want to acknowledge my thesis advisors John Tsang and Ken Smith for their outstanding support. John was always willing to discuss and inspire such creative ideas and had patience in letting me learn by making many mistakes. Most importantly, his existential outlook deriving meaning from the process of science itself, outside of any external expectations helped me grow as a scientist and as a person. Working with Ken showed me the value of truly encyclopedic knowledge and an ability to quickly discern the aspects of scientific inquiry that matter the most to patients. I aspire to this skill and to Ken's ability to create a sense of community in the lab and institute.

I also owe special thanks to Andrew Martins. The early days optimizing single cell experiments and "seeing" the immune system in a new way for the first time together were among the most exciting times I've had in science. I was very fortunate to work closely with such a rigorous scientist, kind mentor, and friend.

From Cambridge, over the years Paul Lyons has always been open to helping with the most detailed aspects of projects and I always enjoy our far-reaching conversations. I owe a great deal of thanks to people that came before me in Ken's lab in particular Laura Messer and Victoria Clarke for their organizational superpowers, Santi Kotagiri, Laura Bergamaschi, and Aimee Hanson for their massive efforts on projects form the basis of ongoing exciting work that is not in this thesis. Thanks to Eoin McKinney and Gosia Trynka for critiques of my first year report and Eoin for the continued conversations. The kindness of Katrin, Johanna, Dan and everyone from the Smith lab made my experience in Cambridge a happy and memorable time.

From NIH, Thanks to Pam Schwartzberg for insights in interpreting results on multiple projects and Chen Zhao for initiating the irAE study and providing career advice. Thank you to the Oxcam office and the International Biomedical Research Alliance for their support. Many people from CHI and the Tsang lab deserve acknowledgement. Thanks in particular to Thorsten, Mani and Julián for sparking my interest in systems biology / data science years ago, Foo, Yuri, and Rohit for their computational expertise, Neha and Jinguo for helping with CITE-seq optimization, and for the many conversations over the years, thanks to Kyemyung, Nick, Can, Rachel, Darius, Dylan, John Kim, Laura and William. I'll miss our meandering lab meetings.

Thanks to Chris Hourigan for his outstanding mentorship and for inspiring my journey on the physician scientist path. Thanks to Andy Camilli and Mara Shainheit for encouraging me as an undergraduate, and Harry Bernheim for his incredible immunology classes. Special thanks to Lori Koziol for giving me my first opportunity to work in a lab and encouraging me to continue a research career. Thank you to Dan Hollern, friends from UNC, and the MSTP who supported my journey to NIH/UK.

The pandemic obviously created lots of uncertainty–thanks to Oxcam / Katie Stagliano for her thoughtful support at this time, as well as Tom and Amelia for their hospitality, and Bean and Lottie for their emotional support.

Enormous thanks are owed to my parents Bob and Lisa for giving me an incredible amount of freedom to choose any path in life. They provided me with unwavering support me no matter how outlandish my ideas and goals seemed at times. Thank you to my sister Mandy for lifelong support and for giving me the great honor of my first invited academic lecture to her third grade class.

Finally, to my wife Sarah: Thank you for your patience and above all, for your friendship. The hardest parts of graduate school were the most meaningful because I spent them with you.

Contributions

Chapter 2: I performed CITE-seq experiments together with Andrew Martins. We created the dsb method together with John Tsang. I created the dsb R package and performed the analysis in the manuscript.

Chapter 3: CITE-seq experiments were done by myself and Andrew Martins. The CyTOF experiments were done by Brian Sellers and Juan A. Quiel. Foo Cheung assisted with genotype data for sample demultiplexing. Rohit Farmer created the HDSTiM package and assisted with CyTOF data processing. I created the other computational methods and performed the analysis in the manuscript.

Chapter 4: I performed CITE-seq experiments together with Andrew Martins who also performed the alignment and cell clustering. Chen Zhao secured funding for and supervised the project with John Tsang. I created the computational methods and performed the analysis in the manuscript.

Work done during my PhD not included in this thesis includes the following:

The paper below was led by Yuri Kotliarov and Rachel Sparks and referenced in the introduction. I generated the CITE-seq data in the paper with Andrew Martins. I did analysis using early versions of the computational methods described in chapters 2 and 3 which contributed to parts of Figures 3–6.

Kotliarov 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 DOI: 10.1038/s41591-020-0769-8

Three papers below were done in collaboration with the Human Immunology Project Consortium. I did quality control work, analysis, and interpreted results at weekly meetings.

Diray-Arce et al. The Immune Signatures Data Resource: A compendium of systems vaccinology datasets. *BioRxiv* (2021) (*in press*) DOI: 10.1101/2021.11.05.465336

An innate immune activation state prior to vaccination predicts responsiveness to multiple vaccines. *BioRxiv* (2021) (*in press*) DOI: 10.1101/2021.09.26.461847

Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses. *BioRxiv* (2022) DOI: 10.1101/2022.04.20.488939

I contributed analysis software and helped with analysis of this work on severe COVID-19:

Liu et. al. Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* (2021) DOI: 10.1016/j.cell.2021.02.018

I designed single cell experiments, helped with analysis and trained researchers to do bioinformatics and analysis for this paper evaluating a vaccine platform.

F Baharom et al. Intravenous nanoparticle vaccination generates stem-like TCF1+ neoantigen-specific CD8+ T cells. *Nature Immunology* (2021) DOI: 10.1038/s41590-020-00810-3

Reproducibility Statement

The code to reproduce all of the analysis and recreate all figures in this thesis can be found at the following repositories:

Chapter 2: https://github.com/niaid/dsb_manuscript

Software package associated with Chapter 2: https://CRAN.R-project.org/package=dsb

Chapter 3: https://github.com/niaid/fsc (not yet public as of Aug 2022)

Chapter 4: https://github.com/niaid/irae_manuscript

CD – Cluster of differentiation

LPS – lipopolysaccharide

ISG – Interferon Stimulated Gene

IFN – Interferon

TNF – Tumor Necrosis Factor

PRR – Pattern Recognition Receptor

PAMP – Pathogen-Associated Molecular Pattern

TLR – Toll-Like Receptor

cQTL – Cytokine Quantitative Trait Loci

eQTL – Expression Quantitative Trait Loci

GWAS – Genome Wide Association Study

CITE-seq – Cellular Indexing of Transcriptomes and Epitopes by sequencing

CyTOF – Cytometry by Time of Flight

Tfh – T-Follicular Helper Cell

NK – Natural Killer Cell

mDC – Myeloid derived Dendritic Cell

pDC – Plasmacytoid Dendritic Cell

TIV – Trivalent Inactivated Influenza Vaccine

LAIV – Live Attenuated Influenza Vaccine

YF – Yellow Fever

ICI – Immune Checkpoint Inhibitor

irAE – Immune Related Adverse Event

TABLE OF CONTENTS

LIST OF FIGURES

1 TOP DOWN HUMAN SYSTEMS IMMUNOLOGY WITH SINGLE CELL MULTIOMICS

1.1 Introduction

Biological systems are organized in a hierarchy of scales from molecules and cells to organisms and populations. Higher level biological system behaviors, such as the formation of immune memory, are "emergent"; the coordinated activity of molecules, interacting cells, and functional changes within tissues cascade into these higher level system behaviors. Cell and molecular biology played substantial roles in immunology research of the past 7 decades in confirming theories of clonal selection[1], defining molecular mechanisms of immune memory[2,3] and clarifying interconnectedness of innate adaptive arms of the immune system[4–6]. Genomic technologies of the past two decades have brought the complexity of these lower level immune system properties further into focus. Since the identification of just two specialized lymphocyte subsets more than 50 years ago[7] we now recognize an array of specialized immune cell subsets with distinct protein[8,9], chromatin[10,11], epigenome[12], transcriptional[13–16] and cytokine circuitry.

Integrative systems approaches applied to understand the logic of gene regulatory circuits and cell signaling components have long had a role in molecular biology and physiology[17]. More recently, advances in high throughput technologies have empowered full system scale statistical modeling with "top down" approaches. Without the catalogues of system components and their lower level interactions identified by reductionist studies, the top down approach would have no scaffold on which build a quantitative understanding human phenotypic variations. However, reductionist

approaches using model organisms have limited capacity to directly explore unbiasedly why identical immune system perturbations applied to the human population result in variations in the emergent response[18]. These outcome variations in the population are readily apparent in SARS-CoV2 infection response[19], vaccination response[20] and immune checkpoint inhibition[21,22]; all topics explored during this thesis work. Human immune systems are shaped by unique experiences–the antigen ecology encountered in an individual's environment[23–27], intrinsic factors including age, sex, and genetics, and the interaction of these variations with nonlinear and stochastic processes governing cellular response[28].

Top down approaches use vaccination[29,30] as an ethical *in vivo* perturbation, or *ex vivo* stimulation of cells[31] collected from human cohorts followed by quantitative modeling of molecular responses to develop a quantitative and predictive understanding of human immunology[32,33]. Linking molecular immune system variations to emergent response variations requires rigorous analytical frameworks[20]. In this work, we develop approaches to integrate immune states across biological scales to explore how multicellular immune system states may cascade into different emergent response outcomes.

Given the complexity and multiscale nature of immune response variation, it can be helpful to group immune system organization into different biological layers. We next provide an organizing framework for conceptualizing layers of immune system information, review the experimental methods that profile these layers at distinct phases of the response, and describe the types of data created form these technologies in the context of human cohort studies. An overview of challenges in computational/statistical approaches for linking molecular and emergent responses is provided to motivate our adoption of new multimodal single cell sequencing technologies. We then describe challenges in implementing these new technologies in order to frame the tools and analysis frameworks developed in this thesis aimed at addressing them. Finally, we provide a comprehensive review of insights emerging from top down systems immunology studies that use this *ex vivo* and *in vitro* perturbation. From this synthesis, we describe the idea of immune "setpoints" which have emerged from our own work and other studies finding the baseline immune system state can influence molecular and emergent responses. We further build on these ideas in the results chapters.

1.2 Experimental and analytical approaches for top down human systems immunology

1.2.1 Layers of immune system organization

The composition of the human immune system exhibits marked variation[34]; frequency of cells in blood provides a high-level organizing layer of immune system composition. Cell frequencies have dynamic patterns; at the shortest time scales, cortisol controls circadian immune cell egress and tissue homing[35] which makes effector T cell populations peak at night, while naive and memory subsets peak during the day[36]. Seasonal changes in cell frequency have been inferred from transcriptomics[37] although high resolution measurements showed lack of seasonal effects in more finely resolved subpopulations[38]. Besides these rhythmic variations, on the scale of months, person to person variation was higher than within individual longitudinal variation, in general across 126 cell subsets in a cohort of over 60 individuals, adjusting for age, sex and ethnicity[39]. The same pattern was later observed in a geographically distinct (Belgian) population of 177 individuals[40]. More recently a study of 99 healthy adults age 50-65 found elevated within-individual longitudinal cell frequency variation was associated with poor metabolic health[38]. Cell frequency provides just one organizing layer of immune system information exhibiting substantial between individual variation[34]. Below we group layers of immune system organization defined by both biological information and the types of data that can be collected and integrated with other layers in multiscale analysis.

Layer 1: human population scalar measurements. Layer 1 data have one scalar measurement per individual that cannot be further deconstructed, for example, an individual's age. Response quality surrogates are often be measured at this scale including the fold change in antibody titers following vaccination. Population variation in these quality measures can be modeled as a function of other layer 1 data, or any layer below.

Layer 2a: population variations in immune cell frequency or number Layer 2a is the frequency / counts of immune cell subsets across individuals. Cell frequency data are distinct from layers below because they exist at a distinct cellular level of biological organization (as opposed to molecules). Cell enumeration is often reported as frequency relative to a parent population, though absolute counts can reveal lymphopenia when

relative lymphocyte frequencies are similar[41]. These data form a 2-dimensional matrix of cell types by individuals, often with repeated measures over time.

Layer 2b: molecular measurements at the individual level Layer 2b has the same structure as layer 2a, but describes a lower biological level of organization in the form of molecules across individuals. There are 2 classes of layer 2b data. “Native” data are linked to a single biological source and can’t be further broken down, such as the plasma proteome. “Aggregate” data are derived from a heterogeneous background, for example, bulk tissue gene expression measurements–these can only be further deconvolved to cellular constituents computationally.

Layer 3: Molecular phenotypes within immune cell populations. Layer 3 are molecular measurements like layer 2b but are further indexed within distinct immune cell subsets which pinpoint information masked by bulk profiling[42]. Layer 3 data include chromatin accessibility or gene expression within subsets and are often derived from cell sorting followed by sequencing. This data can be aggregated as a single matrix though there can be advantages to using a 3-dimensional data structure, where features are separately filtered and modeled for each cell type.

Layer 4: Molecular phenotypes within single cells. Layer 4 data are molecular measurements of single cells and multiomic single cell methods[43–47]. The structure of these data includes multiple sets of features (e.g. genes and proteins) for all cells belonging to each individual. Importantly, multimodal protein and mRNA data can be re-aggregated to reconstruct the biological layers above. Findings from higher layers can be deconstructed with these data down to the level of single cells. Layer 4 data can be used to create computational reconstructions of cell states on lower dimensional embeddings[48–50] which can also be used to interpret statistical results from higher layers (Fig. 1.1).

1.2.2 Experimental technologies to capture immune response phases at different organizing layers

The purpose of perturbing a complex biological system is to identify regulators and effectors in order to build models that can make predictions and ultimately infer the manipulations that elicit a desired change[33,51]. Systems biology studies of vaccination attempt to move toward the latter goal by mapping correlates of protection to other lower layer data using mathematical and statistical machine learning models.

Perturbations in these studies generally take two forms: 1. *ex vivo* stimulating blood immune cells with a pathogen or cytokine and measuring effector responses[31]. 2. *in vivo* vacination[20]. Different experimental technologies capture different phases of the multicellular orchestrated immune response. Cell stimulation approaches offer a focused look at one of the phases outlined below, whereas vaccination invokes the entire synergistic response process.

Immediate response phase: The immediate phase reflects the initial cell-intrinsic signaling cascade in response to antigen or cytokine stimulation on the scale of minutes. Common stimulants include lymphocyte receptor stimulation[52,53], whole antigen (heat killed or live pathogens), and partial antigens, including pathogen associated molecular patterns (PAMPs) like lipopolysaccharide (LPS). After stimulating specific pattern recognition receptors (PRRs), second messengers relay phosphoprotein signals which can be measured optimally 15-30 minutes after stimulating cells using ELISA (creating layer 1 data) or intracellular flow cytometry and Cy-TOF for single cell measurements[54,55] (layer 4 data). A hallmark of the immediate response phase is the activation of what have been defined as “class B” transcription factors[56,57]. These mediate the “immediate” response because they are present in the cell in inactive form, for example CREB, HIF-1a, NF-kB, and STATs[56,57]. Population variations in these phosphosignaling relay proteins thus represent variations in cell-intrinsic signaling capacity to regulate the transcription the phases of the response below.

Early response phase Activation of pre synthesized transcription factors above mediate transcription of early “primary response genes” (PRGs)[56,57]; immune responses are mediated by combinatorial induction of broadly active and cell type specific PRGs, which include other transcription factors mediating further induction of secondary response genes[58]. This middle phase can thus be broken down into several sub-phases controlled by “Class C” transcription factors inducing either broad (C1) or more cell type specific genes (C2 and C3)[56]. This coordination is governed by structural aspects of the epigenome[59] which result in coregulation of response genes with lineage-specific transcription factors such as SPI1 in monocytes / macrophages[58,60]. Distinct classes of genes are induced within this phase lasting from hours to the early days after stimulation. Metabolic changes and secreted cytokines can also be profiled from this phase. The early response phase can be profiled with bulk or single cell gene expression measurement up to around 4-5 days following ex vivo stimulation or in vivo

vaccination. At these timescales, it can be difficult to resolve the induction of specific early genes from their downstream regulated targets such, as those regulated by both type I and type II interferons.

Late response phase: The early response genes induce soluble or physical communication relays between other cells and tissues. For example, interacting cells from blood and lymph organize into germinal centers hours to days following vaccination. These cell dynamics and interactions can be profiled from blood 1 to 4 weeks following stimulation (most relevant to vaccination studies), where expression changes likely reflect cellular dynamics. Serum measurements can also profile the circulating cytokines mediating cell-cell communication relays at this phase.

Emergent response phase Measures of end point response quality make up the "emergent" response dictated by the dynamic processes in the earlier phases. For example, after vaccination, generation of antibody producing B cells proceeds through interactions including T follicular helper (tFH) cells[61] memory and naive B cells, and other antigen presenting cells[62]. The functional response can be assayed weeks to months following vaccination through measurement of IFN-g release from cytotoxic T cells, or the fold change in antibody titers.

1.2.3 Technologies and analytical frameworks to identify multicellular immune networks linked to outcomes

Top down human studies aim to define molecular features which are 1. linked to individual "intrinsic" factors including age sex and genetics; 2. Induced by perturbation coherently across individuals; 3. Correlated with emergent response quality such as post vaccination antibody titer fold change. Population variations and longitudinal profiling can identify robust correlates of response outcomes in top down vaccination studies[20]. Below we refine these ideas for cohort studies on single cell multiomics and describe the motivation for the analysis framework developed for this thesis.

Bottom up approaches using ordinary differential equations to model system dynamics usually assume prior knowledge of connections between components of a closed system. Top down methods instead rely on perturbation and measurement of the system *en masse* to define the impacted features and link them to population variations in higher level behaviors. To accomplish this, most studies reviewed in the sections below

utilize generalizations of multivariate generalized linear models (GLMs) including machine learning methods which internally depend on GLMs.

These approaches model the deviation of each feature from the mean of sampled individuals as an additive combination of modeled covariates. Coefficient estimates from the population samples may represent a thin slice of the environmental landscapes which can interact with covariates to modify the level of the modeled feature[63]. Scaling these studies to thousands of diverse individuals in the future will help alleviate this sampling bias. The effect size or fraction variance explained by each intrinsic covariate can then be calculated for each feature (i.e. gene and protein). Collinearity (correlation) among covariates themselves (e.g. intrinsic variables) can make interpretation of their individual effects challenging. Two correlated variables can also relate to the feature being modeled in a causal structure where one variable mediates the effects of another on the outcome. Directed acyclic graphs and structural equation models can help identify and quantify these effects[64–66].

Data returned by top down technologies commonly include sequencing-based outputs which yield molecule counts. Dedicated statistical count models exist for bulk(layer 2 and 3)[67,68] and single cell data[69]. A limitation of the dedicated statistical software for analysis of sequencing counts is the relative lack of available methods for multi sample time series data with repeated measurements from each individual. For example, software for implementing these count models do not have available functionality to estimate covariance between repeated measures from the same individual with varying / random effects. The Limma software[70] uses a single genome wide average covariance applied to all features, however Hoffman and colleagues[71] showed how this approach can inflate error due to systematically under or over estimating the intra subject correlation for features below or above this average. The authors then extended multilevel linear models to RNAseq data for complex experiment designs. The approach uses efficiently parallelized functions for maximum likelihood estimation with per gene covariance, enables arbitrarily complex random effects structures using lme4[72], and integrates observation level weights accounting for mean variance count trend[73]. We adopted this method and implemented for subset level modeling (layer 3) and further deconvolved results using additional single cell multilevel models and computational reconstructions of cell state, as further detailed in the methods sections of chapter 3 and 4 (see fig. 1.1).

Top down approaches with highly multiplexed measurements are "p > n" problems[74] with more modeled features than samples, creating challenges both for building predictive models and type I error inflation. Several approaches exist to help limit false discoveries; in addition, the high modularity of biological processes[75] can be leveraged to take advantage of this dimensionality challenge. At the individual feature level, Empirical Bayes methods help regularize coefficient estimates toward the genome wide trend[76]. Monte Carlo cross procedures[20,39] and various machine learning methods can also define only the features most robustly linked to outcomes through use of regularizing priors[77,78]. Some studies set out to define a minimal set of robustly "differentially expressed" features passing a false discovery cutoffs. When combined in a meta-analysis framework, this has a proven valuable for developing clinical stratification tools[79,80]. However, in many studies outlined below, this approach is also used to disentangle and understand the biology of the perturbation. Eliminating all but the strongest of "differentially expressed" features from consideration can be limiting because this ignores known biology of the data generation process in favor of dissecting the biology of a class of genes defined by statistical properties (large effect size and small variance). Biological systems are highly modular[75] with structured regulatory networks[81,82] (e.g. see "early response phase" above). For this reason, computational approaches that attempt to 1. reconstruct gene regulatory networks directly[83–85], 2. group genes into functional modules via data mining or mutual information[86–88], or 3. unbiasedly construct correlation networks[89,90], offer an opportunity improve identification of coherent and predictive biological processes induced by perturbation. Importantly, these approaches often identify biological processes that would not meet statistical cutoffs in gene-centric analysis. A related process for recovering modular patterns is to compress data in unsupervised fashion into lower dimensional vectors which capture the maximal variation across features (PCA) or in a way which simultaneously captures variation of an outcome variable (e.g. partial least squares). These compressed dimensions can then be interpreted or used to form network structures for further statistical or mathematical analysis such as topological data analysis[91].

1.2.3.1 multi-modal single cell technologies for multiscale modeling

Determining whether observing a gene or pathway after perturbation represents an intracellular transcriptome state transition vs changes to cell type composition is a major challenge. Some studies simply use transcriptome measurements as a proxy for cell frequency[92] through computational deconvolution algorithms[93]. The recent development and commercialization of droplet microfluidic technologies[94–96], has led to massive uptake of single cell RNA sequencing technology[97]. A second wave of technology improved both the scalability of scRNAseq through sample multiplexing[98,99] and the development of multi-omic data modalities[43] including simultaneous protein phenotyping[44,45] such as CITE-seq (Cellular Indexing of Transcriptomes and Epitopes). In initial pilot experiments we found these techniques to be promising, however there was abundant noise in protein data which obscured the identification of true cell populations. We address this in chapter 2 through experimental deconvolution of protein noise sources and development of a dedicated method, dsb, for protein normalization[100].

Best practice computational analysis tools for clustering, integration, cell type identification and downstream analysis of single cell data[101,102] are not well suited for complex human cohort experiment designs including repeated timed pre-post perturbation measures from individuals nested in different outcome or treatment groups. To address these challenges, we created computational pipelines to implement multilevel models within immune subsets defined by dsb normalized protein level. These methods created maps of highly interpretable immune cell perturbation phenotypes coherently induced across individuals and varying between outcome groups. We then further deconstructed these robust statistical results by understanding feature behavior along lower dimensional embeddings which position cells along a continuum of intermediate states[48–50,103]; for example combining these pseudotime methods with real time kinetics relative to perturbations and protein data helped identify processes such as differentiation obscured in aggregate data. We further interpreted results by leveraging population variations to create correlation networks from these multicellular baseline and perturbation phenotypes. Together these approaches create a framework for quantifying multicellular networks linked to coherent and predictive immune responses (Fig. 1.1).

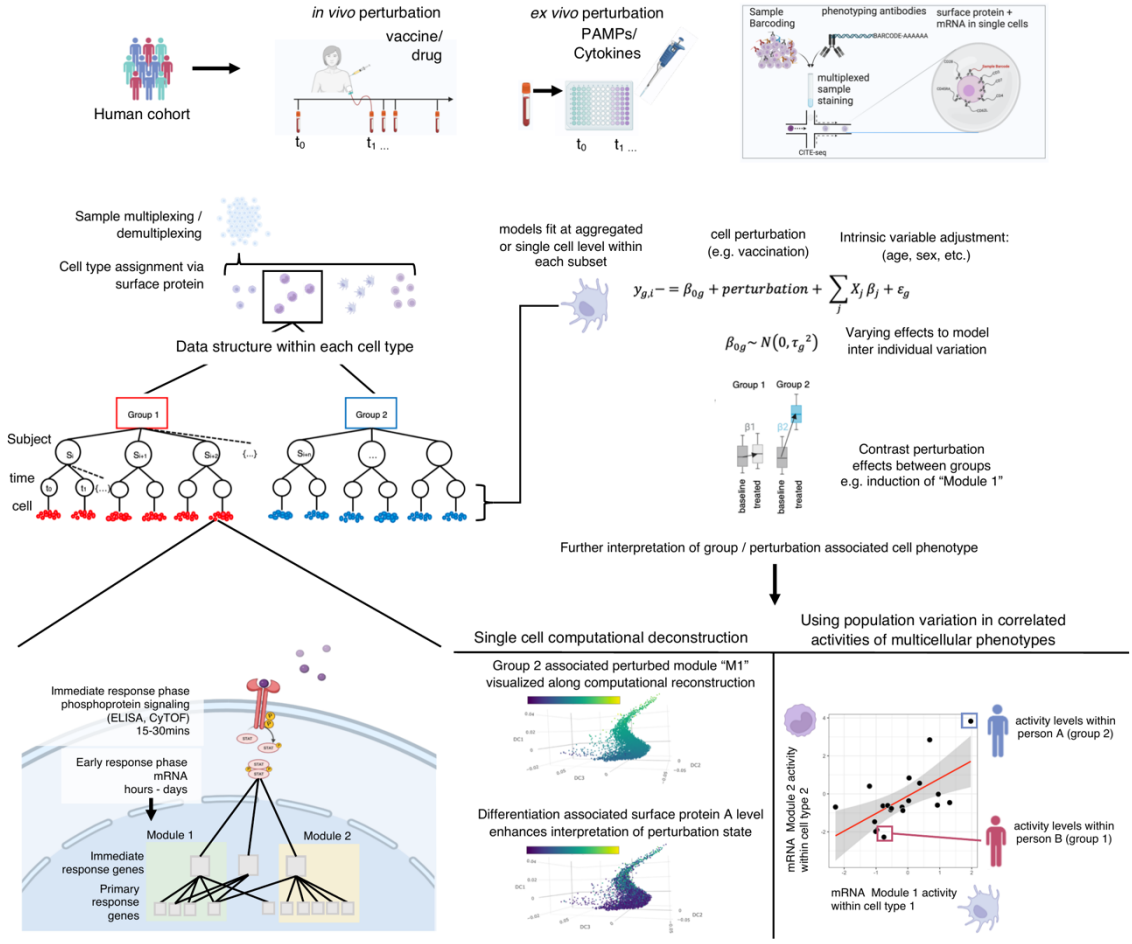


Figure 1.1 Multiscale analysis for top down single cell human immunology.

The Structure of the data shown are a theoretical top down systems immunology study with the same structure as the data in this thesis work. "Layer 4" single cell molecular data e.g. CITE-seq and CyTOF are collected to deconvolve baseline and pre vs post perturbation immediate (assessed by phosphoprotein level after *ex vivo* stimulation) and early / late (assessed by single cell mRNA after *in vivo* vaccination) timed perturbation response phenotypes between two or more groups with different emergent outcomes, e.g. antibody response at a later (d28 or d70) timepoint. At the single cell level, the immediate response phase is shown as phosphorylation of a STAT protein (bottom left); CyTOF measurement assesses phosphoprotein levels at single cell resolution in combination with phenotyping antibodies to pinpoint immediate signaling phenotypes. The early response phase is broken down into functional modules based on co-expressed mRNA "modules" aggregating genes using prior knowledge or based on correlated expression. Both module expression and phosphoprotein expression assessed with CITE-seq and CyTOF can be aggregated to the mean activity of "layer 3" cell subsets based on surface protein phenotypes. A multiscale modeling approach for CITE-seq analysis is shown on the right where the timed perturbation phenotype of a gene module "M1" is modeled with a mixed effects generalized linear model incorporating population intrinsic variables (e.g. age and sex), the perturbation effect, and a varying / random effect to model variation in baseline expression across donors. Statistical contrasts applied to the model fit can identify perturbation phenotypes differing between groups after adjusting for these effects, as shown by the box plot distribution of "M1" across the two response groups. The identity of M1 can be further

understood by visualizing its expression in a computational reconstruction of single cell states (bottom middle). In this case expression of M1 increases with increasing values of component 1 on the y axis. The reconstruction can be further interpreted by incorporating a different data modality collected from the same cells. For example, in this case, a protein associated with differentiation, "Protein A", increases along component 1 as well. Putting all of these results together, we can interpret "M1" as a dynamic differentiation process perturbed coherently by the treatment after adjusting for age and sex specifically in group 2 but not group 1. When we carry out this analysis on all cell subsets on many modules, we can use the population variations in the cohort to examine their correlated activities to understand circuitries of cells differing between groups or perturbed by treatment (bottom right). This reveals, for example, that M1 is correlated across individuals with a different process M2 in a different cell type.

1.3 Insights from systems biology studies of human immunology

We next highlight themes and concepts which have emerged from studies taking top down systems biology approaches, using ex vivo cell stimulation and in vivo vaccination, to study human immunology. Several challenges make comparison of these studies difficult. In general results are interpreted / reported nonuniformly, as expected for an emerging field emphasizing innovation. For example, for transcriptome analysis, some authors take a "differential expression" gene-centric approach instead of interpreting co-expressed gene modules, others use models without adjustment for covariates, and many profile relatively small cohorts. In addition, studies commonly rely on qualitative comparison of separate analysis of sub groups. This last point in particular is important; for example, if a perturbation induces a change in one group with adjusted p less than 0.05 and another group with adjusted p greater than 0.05, this is not equivalent specifically testing whether the *difference* in in fold change between the groups is "significant" at that same $p<0.05$ level[104]. To test the differences in perturbation effects between *a priori* defined groups, we emphasize the use of statistical contrasts throughout the results chapters to incorporate information across groups. Lack of adjustment for covariates is also important as their effects can be considerably linked to molecular features and obscure identification of "root" correlates of response[20]. With these caveats in mind, below we aim to integrate information gleaned from these studies, comment on major emerging themes and highlight areas for further study.

1.4 Genetics

1.4.1.1 Genetic control of the human immune system

During the Neolithic rise of animal domestication, zoonotic transmission expanded human pathogen burden leading to new diseases in population dense areas[105]. It would be expected that variants impacting the ability of immune cells to sense the landscape of pathogens would be critical for adaptation to the recent shifting pathogen landscape. Indeed, evidence exists for purifying selection against pattern recognition receptor loci[106] that can cause immunodeficiency[107,108], while balancing selection maintains allelic diversity at HLA class I[109] and interleukin gene loci[110] in geographic regions with high pathogen diversity.

Studies of immune-mediated disease have also shed light on genetic control of immune function including single gene underpinnings of sporadic non-familial primary immunedeficiency[111] and complex inheritance of common immune mediated diseases[112]. Within-cases GWAS have further defined how variants linked to disease prognosis are distinct from disease risk loci in inflammatory bowel disease[113]. These studies have defined how genetic variations can lead to diverging immune response behaviors depending on context. For example, an allele linked to better Crohn's disease prognosis via reduced inflammation increased susceptibility to malaria in Kenyan and Vietnamese cohorts[114]. The genetic architecture of immune function can thus represent a two edged sword depending on interactions with an individual's environment. Our findings on shared immune system setpoints prior to trigger in healthy individuals and lupus patients associated with beneficial or detrimental plasmablast activity respectively, mirror this finding[115] (see Fig. 1.2 below).

Globally, genetics tend to impact a smaller number of molecular phenotypes than other intrinsic variables, but with greater effect sizes[66]. Combinatorial flow cytometry panels have defined GWAS loci linked to variable immune frequency[9] and genetic effects have been shown to have relatively larger effects on innate cell frequency[116]. Genetics can also control "layer 4" variations through association with cell to cell surface protein dispersion within an individual[117]. The genome also contributes to more fine-grained variations in disease traits as demonstrated by recent statistical analysis of population genetics data. Modern GWAS hits explain much less than predicted genetic variation, "missing heritability"[118] later shown to be due to aggregated effects of common variants with less than "genome wide significant" effects[119,120] enriched in non-coding

regions[121]. One study identified that while GWAS hits for complex traits (height) have higher *individual* effect sizes, common variants are massively enriched for positive effect sizes with low p values which contribute a majority of total heritability and are uniformly distributed across the genome in regulatory regions[122]. The authors proposed most expressed genes are highly connected through regulatory networks to the nearest "core" GWAS hit, explaining why most heritability derives from genes with less obvious effects[122]. These results implicate comprehensive mapping cell-specific gene regulatory architecture induced by immune perturbation as a crucial element to truly understand human immune variations in health and disease.

1.4.1.2 Genetic control of immune ex vivo cell stimulation phenotypes

The impact of genetic variation on cytokine response was assessed in more than 400 individuals after stimulation of PBMCs using models for age sex and cell counts[123]. Cytokine qualitative trait loci (cQTL) were enriched for PRR genes and their myeloid specific enhancers in a context dependent manner and significantly correlated with more than 20 cytokine stimulation responses. These loci overlapped with regions undergoing positive selection[124] and the same variants increased risk for infections, demonstrating evolution tuning immediate phase immune responses. Another report of 197 Europeans used a similar approach to compare bacterial and fungal stimulation of PBMC[125] to investigate genetic control of *ex vivo* cytokine responses. GWAS traits binned into functional categories (e.g. metabolic related, cancer related) identified infectious disease, heart disease, and immune-mediated disease enriched for cytokine quantitative trait loci (cQTLs). Infectious disease susceptibility loci were associated with lower cytokine production, while immune mediated disease were linked to higher cytokine production. Correlated response fold change across individuals demonstrated population variation in cytokine response levels were stimulation-specific rather than cell type specific. This suggests context specific genotype effects, rather than global genetic control of cytokine production. Interestingly cytokines thought to be derived from T helper 17 cells (TH17) which have roles in autoimmunity[126] and bacterial/fungal response[127] formed a separate, pathogen-independent correlated cluster indicating certain individuals have elevated TH17 response potential.

Genetic control of interferon circuitry was also identified in a report of monocytes stimulated with LPS and interferon from 432 Europeans[128]. Expression quantitative trait loci (eQTL) pairs identified were linked to time-dependent downstream effects, for

example a variant associated with lower interferon beta receptor expression following 2h of LPS stimulation propagated to downstream ISGs in cis at 24h. Another study of innate cell genome architecture stimulated dendritic cells (DCs) from African, Caucasian, and East Asian descended individuals with LPS, influenza and IFNb[129]. This study found 57 variants linked to all three stimulation conditions, including LPS, that reflected a shared induction of the interferon pathway. These results support the notion that evolution tolerates significant variation in interferon circuitry in health, as is the case in disease states[130].

In a study of transcriptional responses to CD4 T cell stimulation[53] a positive correlation was observed between tonic CD4 T cell IFN transcripts and their levels following 48h of TCR stimulation with TH17 polarizing cytokines, suggesting tonic IFN response genes may control a positive feedback mechanism relating interferon dependent inflammatory cytokines. Stimulation dependent variants implicated in autoimmune disease risk[131,132] were also linked to IL2 signalling (IL2RA). The authors hypothesized protective alleles localized to regulatory CD4 T cells (Treg) which echo more recent mouse studies of self-activated T cells found to be locally constrained via Treg activation through their IL2R[133]. Another study of mono (MZ) and dizygotic (DZ) twins aged 8 to 82 years measured cell frequencies, serum proteins and stimulated 8 cell populations with cytokines to measure the immediate response within single cells[25]. Structural equation models revealed phosphoprotein response to most cytokine stimulations had little evidence of heritability, yet a small number had substantial genetic effect sizes, a pattern seen in other comprehensive studies reviewed below[66]. pSTAT5 response to IL2 in T cells were among the most heritable responses, supporting a role for genetic control of immediate T cell signalling circuits that control the early response genes identified through transcriptomic analysis above[53].

1.4.1.3 Genetic control of timed vaccine perturbation response

Twin studies have demonstrated no heritability in antibody response to TIV vaccination[25]. To identify more fine scale genomic regulators of influenza vaccine response, Franco and colleagues genotyped two cohorts of over 100 individuals and measured gene expression at baseline, and day 1, 3 and 7 post vaccination[134]. The authors used mixed effects models and a conditional independence test framework[135] to infer causal (loci which modify gene which modify antibody response) SNP transcript pairs with changes in variance significantly explained by genotype over different

timepoints post vaccination. Genetic effects at multiple loci were only apparent after the vaccine perturbation and gene-transcript pairs with high variance were enriched for antigen presentation, cytotoxic T cell activity, DC maturation and membrane trafficking. Antigen presentation genes were most strongly coupled to humoral immune response and trended towards being causal though the study was underpowered to define these effects. Overall the study points to genetic variation in antigen presentation capacity apparent only after timed *in vivo* perturbation linked to improved vaccination response.

1.5 Sex

1.5.1.1 Sex effects on immune system function

Sex differences in immune system function can be attributed to both increased X chromosome immune gene expression and hormonal control of immune signalling[136]. Androgens (testosterone) suppress immune cell activity through increasing expression of anti-inflammatory cytokines[137]. Conversely, estrogens can stimulate monocyte and lymphocyte activation and differentiation and can increase expression of proinflammatory cytokines[138]. Incomplete X inactivation can also play a role in mediating increased responses to immune cell stimulation, including (X chromosome) TLR7 induced interferon response in pDCs[139]. These may influence vaccination responses, as antibody titers after influenza vaccination are typically higher in women than men, though this varies by vaccine[140].

1.5.1.2 Sex effects on ex vivo stimulation phenotypes

Sex effects were well characterized in a study which stimulated cells from an age, sex, and ethnicity balanced cohort of 1000 individuals with a host of pathogens[66]. Of more than 500 sex dependent genes, 181 were only mediated by sex only after stimulation; of those, 45% were sex dependent after all 6 stimulation conditions. The analysis framework used by the authors incorporated age, blood cell composition and technical variables and structural equation models. Increased CD4 T cell and reduced monocyte frequencies in women mediated some of the sex-dependent effects. The mediation of gene expression changes through monocyte frequency were consistent with cytokine secretion in response to pathogen stimulation in a different study, which found men had higher monocyte derived serum cytokine responses to most pathogens[141]. This study attempted to define roles of sex steroids by sub group analysis of women taking oral

contraceptives which identified lower IFN-gamma and TNF-alpha after LPS stimulation in the oral contraceptive group. These steroid hormone effects appear to be relevant in humans *in vivo* as outlined below.

1.5.1.3 Sex effects in systems biology studies using vaccination

Sex mediated transcriptional responses have also been linked to post vaccination antibody levels, with a role for both sex effects on B and T cell activation[142]. Two early top down studies of yellow fever vaccination responses did not adjust for intrinsic covariates in estimating vaccine effects on the transcriptome[143,144]. Klein and colleagues reanalyzed these data using 2 way analysis of variance to estimate effects of vaccination and sex[136] and found 10-fold more genes passing an FDR cutoff in females compared to males as a function of time post vaccination, including many interferon response genes (ISGs). This reanalysis did not test marginal effects at certain timepoints or adjust for other variables (e.g. age), nonetheless, they demonstrate the importance of consideration of population covariates in human studies.

A later report found roles for hormonal control of immune states in analysis of sex effects on influenza vaccination responses[145]. The authors found an interaction effect predictive of antibody responses between sex and the baseline expression of a module of genes related to lipid synthesis. Modeling antibody titer fold change as a function of the interaction between sex and the lipid module revealed opposite effects in males and females, with higher lipid module expression correlating with reduced antibody responses in males (log odds < 1) but increased responses (log odds > 2) in females. Stratifying men by testosterone level showed the interaction between testosterone level and the lipid module was only significant for high testosterone males; i.e. the lipid module did not impact antibody responses in the male lower testosterone group or females. These thoughtful analyses provide evidence for hormones shaping the baseline immune state in a sex-specific manner, explaining part of the observed sex bias in vaccination response.

1.6 Age

1.6.1.1 age - overview

The immune system gradually converges early in life through a stereotypic developmental trajectory marked by gradually increasing lymphocyte fractions[91] and substantial changes in the thymus[146]. Immune systems then gradually diverge. For

example, in a twin study, heritability of childhood immunizations were found to be high, but adult (influenza) vaccination had no heritability[25]. Evolution appears to favor incomplete negative thymic selection of T cell clones for self-antigens–even HIV- and CMV-specific clones are present in seronegative individuals[147]. It was hypothesized early life infection risk may elicit such a strong evolutionary constraint on reaching sexual maturity, that a more broadly reactive TCR repertoire increases survival probability relative to "opening holes" in the repertoire created by more strict central tolerance[18]. Incidence of autoimmunity rises with age and most genes being positively selected in the human population *increase* risk of development of autoimmune disease[148], lending further support to stronger selection for early life protective immunity with consequences only later in life.

At the cellular level, reduction in circulating Naive CD8 T cells is a universal age related change[149,150]. T cell repertoire changes are reminiscent of synaptic development, with gradual increases then pruning with age[151,152]. Age related reduction in immune response capacity is likely mediated in part by cell intrinsic signalling deficiencies[153]. In addition, cell-to-cell expression variation within individuals can change with age including low distribution of CD38 in T cell subsets, prompting speculation that more uniform CD38 might reflect narrower phenotypic diversity within immune subsets correlating with age[117]. Bulk gene expression profiles indicate inflammatory processes increase with age after adjusting for sex and cell frequencies[154]. Many older individuals develop chronic elevation of blood markers related to inflammation such as CRP, a process thought to involve inflammasome and NFKB activation, increasing risk of cardiovascular disease[155].

1.6.1.2 Age effects on ex vivo stimulation phenotypes

With age, cytokine and transcriptional responses to cell stimulation decline[156]. A comprehensive profiling study measuring single cell phosphoprotein signaling with cy-TOF used partial least squares (PLS) regression to identify combinations of cell frequencies with high covariance with later pSTAT signaling responses to cytokine stimulation. Three latent variables predicted stimulation response to multiple pSTAT signals across many cell subsets, when projected onto this "immunotype" latent space, younger individuals fell within a narrower distribution than older individuals indicating cumulative influence of environmental factors shaped diverging individual

immunotypes over time. Specific factors linked to age included increased memory and effector memory populations, CD161+ CD4+ T cells and NKT cells[157].

In a different report, PBMC cytokine responses of older individuals had globally reduced cytokine production to more than 20 stimulation–cytokine combinations, except for IL1b and IL6 in response to *S. aureus* and *C. albicans* hyphae. Circulating IL1B and IL-6 were also elevated in older individuals at baseline, suggesting amplification of a "poised" age related inflammatory circuit linked to inflammageing[141]. A study of within subject longitudinal stability in immediate phosphosignaling capacity[158] revealed longitudinal stability of older individuals was high for 6 of 17 age associated cytokines but low for the remaining stimulation phenotypes. Older adults tended to have higher longitudinal stability than younger adults, yet had diminished cytokine responses for specific markers with age dependent longitudinal instability. The authors interpreted instability as a hallmark of a degrading cytokine response, mirroring later findings on longitudinal instability linked to poor cardiovascular health[38]. In an analysis of 1000 individuals' PBMC expression responses to diverse stimuli, binning age in 5 levels revealed 20- to 29-year-olds had the strongest transcriptional responses to stimulation with influenza virus (enriched for type I interferon response genes)[66]. Surprisingly, the strongest age difference was relative to that of the 30- to 39-year-old group rather than the oldest individuals. Elevated inflammation seen with age may partially overlap with these ISGs. While these bulk-derived findings have no cellular resolution, intriguingly an *in vivo* study outlined below found similar trends.

1.6.1.3 Age effects in systems biology studies using vaccination

Several efforts have attempted to link age related hyporesponse to vaccination with molecular features in top down studies of vaccination. One report defined age related effects along a continuum of all possible age partitions for the cohort using a "barrier" approach[159] which surprisingly also revealed 30-40 year olds had the lowest relative responses compared to young individuals in 24-h post-vaccination ISG responses. Another study identified a baseline bulk gene expression module enriched for apoptosis genes which covaried with age[160]; multiple predictive models suggested the module was a root correlate of response in the elderly. A later transcriptomic meta-analysis defined baseline correlates of response in older and younger individuals–the interferon and myeloid activation phenotype predictive of robust responses in younger individuals was not predictive in older individuals[161] indicating age-related inflammation signals differ

from the innate / IFN related processes elevated in younger individuals with robust responses. A more recent meta-analysis across 5 vaccine seasons from Yale echoed the findings above; early post vaccination transcriptional response with the most significant meta-effects included innate activation and interferon responses in young individuals which were not correlated with response in older individuals[162] (see Supplementary table 4 in[162]). A similar effect was seen at early timepoints in another metaanalyiss[163] which also noted cell frequency of monocytes in blood increased to the same degree in young and older individuals, indirectly implicating potential cell intrinsic, as opposed to cell-frequency driven transcriptional response effects.

Interestingly, a study specifically looking at the effect of age on vaccination response found stereotypical plasmablast associated gene expression observed on day 7 occurred on day 2 in older individuals, and was absent in frail individuals[164]. This kinetic pattern may thus represent unfavorable processes linked to hyporesponse when seen in other contexts reviewed below. Another report focused on T follicular helper cells which control selection and survival of the germinal center B cells mediating the antibody response[165]. Circulating Tfh (cTfh) cells were found to have an interferon signature lacking in lymph node Tfh cells[i] HA-specific circulating cTfh cells from older individuals did not induce interferon transcriptional response genes that were upregulated in the young, and instead upregulated inflammation associated IL2 and TNF. Mice that could not regulate IL2 were shown to have less Tfh cells and germinal center B cells in the draining lymph node and spleen. This study thus directly links age

[i] Circulating Tfh cells were defined in this study without using CXCR3. The circulating counterpart of Tfh cells were originally defined in a study correlating them with anti-HIV neutralizing antibodies[374]. In that study they were defined as PD-1+CXCR3-CXCR5+. This subset cells had different transcriptional signatures from the CXCR3+ subset; these differences could be an area of further study.

related inflammation to the cells critical for antibody generation in response to influenza.

1.7 Environment

1.7.1.1 Environmental effects on immune system responses in human populations

A major function of the immune system is to sense the environment; antigen experiences accumulate to structure the full complement of lymphocyte receptors into a unique fingerprint of each individual. Age-related divergence in the heritability of early life vaccinations[25], may be due to accumulated environmental exposure causing immune systems to diverge over time. Immune system parameters such as CD4 Treg circulating frequency also exhibit progressively less heritability over time[25]. Reduced cell frequency variation between cohabitating individuals also suggests adaptation to the unique antigen ecology of the domicile[40]. While antigens mediating these adaptations are unknown, it is increasingly appreciated that the microbiome can tune host immune responses[166,167] and may be involved.

CMV may be among the antigen encounters which most strongly perturbs global immune parameters independent of antigen specificity[25]. The repertoire itself appears to also by skewed toward CMV in seropositive individuals[168,169]. The downstream effects of CMV positivity may be associated with cancer and cardiovascular disease[170,171], though molecular mechanisms for these effects are lacking. Additionally there is some evidence CMV positivity increases response capacity to unrelated antigens[172].

Finally, evidence has accumulated for decades that innate cells adapt to local environments. Most recently, there has been a resurgence of interest in non-specific protective effects of vaccination, first described in a 1959 study demonstrating BCG vaccination potentiates subsequent immune responses to tumor challenge[173]. Priming with various stimuli, which could conceivably differ depending on environmental exposure, appears to increase non-specific response capacity through increased hematopoiesis of myeloid progenitors and altered chromatin remodeling around inflammatory and myeloid specific genes[174,175] in mice. Some weaker evidence suggests these same molecular changes may play a role in human innate cell memory as well[176].

1.7.1.2 Environment effects on ex vivo stimulation phenotypes

The effect of CMV seropositivity on responses to non-CMV antigens is difficult to parse due to confounding of CMV seropositivity with age. However by binning

individuals into 9 age groups, it was shown that CMV positivity increased "biological age" in all age bins, even within 20-30 year olds. The biological age was linked to decreased pSTAT signaling in response to stimulation[157]. Cell intrinsic immediate signaling effects were further shown to exhibit environmental effects in the form of seasonality in increased TNF responses to LPS stimulation in whole blood in February and March[177], in line with seasonal elevation of basal inflammation observed in transcriptomics studies[37]. Conflicting data from a large cohort found TNFa gene expression and baseline serum cytokine levels peak in summer[141]. The same study found inferred "monocyte derived" cytokines including TNFa, IL-1B and IL-6 after whole blood stimulation with bacterial, viral and fungal pathogens also all peaked in summer. BMI had a comparatively higher effect on cytokine response to pathogen stimulation of blood cells than sex or smoking status in this study.

In a systematic analysis of healthy Tanzanians, rural compared to urban living was associated with lower inflammatory and interferon basal transcriptional phenotypes. This was unexplained by age, sex, or genetics, but rather was linked to metabolic changes driven by flavones in the rural diet rich in traditional stable cereals, vegetables and banana beer. The flavone rich serum of urban dwelling individuals reversed elevated TNF response phenotypes induced by urban serum stimulation of monocytes[178]. The microbiome likely also mediates effects on immune function through metabolic intermediaries. In work linking the gut microbiome to variations in post stimulation cytokine responses, TNFa levels were positively correlated with fungal stimulations with ~9% of variance explained by 20 microbiome principal components[179]. Metagenomic analysis indicated variation in TNFa and IFNg were related to microbial palmitoleic acid and tryptophan metabolism.

1.7.1.3 Environment effects on molecular responses to vaccination

To test the impact of the microbiome on influenza vaccination responses, an interventional study ablated the microbiome in one group of individuals with broad spectrum antibiotics[180]. After not observing response differences in an initial cohort, the study was repeated using subjects who were naive to flu strains in the vaccine; antibiotic treated individuals had reduced responses to one strain of influenza in this second cohort. Surprisingly typical coherent day 1 and day 7 transcriptional and cellular signatures were observed in both control and antibiotic treated groups, with similar magnitude of effects, perhaps in line with the modest response differences due to

antibiotic treatment. Antibiotic treatment increased expression of AP1 target genes which are linked to inflammation and vaccine hyporesponsiveness in older individuals 24h post vaccination[163]. While this indicated antibiotics temporarily induce an inflammatory program, only secondary bile acid reduction in the treated group was correlated with the lower response to the single strain. The effects of CMV on vaccination responses have been investigated through comparison of young and old individuals with and without CMV seropositivity[181]. Circulating IFN-g levels and CD8+ T cell pSTAT1 and pSTAT3 responses to IL-6 were elevated in the young CMV+ individuals who had higher antibody responses across two cohorts than young CMV- individuals. The authors attempted mechanistic studies with a mouse model of shorter and longer term CMV latency, +/- IFN knockout which supported the idea that interferon effects may mediate the elevated responses in CMV+ young individuals.

A study of the intersection between environment, immune development and vaccine responses defined immune developmental trajectories unified between Dutch and African children from Mozambique and bordering Tanzania. The populations differed in their rate along the developmental trajectory[24], with memory lymphocyte accumulation slower in Dutch children but converging to the level of Tanzanians by age 2. Children from bordering Mozambique also diverged from Tanzanians with increased activated HLADR+CD38+ monocytes, circulating Tfh and plasmablasts at baseline. Children from Mozambique also had increased malaria vaccination (RTS,S) responses compared to Tanzanians. The authors found and experimentally supported the notion that vaccination response differences were due to anemia negatively impacting B cell development in Tanzanians. This study demonstrates how metabolic processes can be mediated by socioeconomic factors influencing immune system development with relevance for vaccination response outcomes.

1.8 Coherent and predictive in vivo immune phenotypes identified in diverse vaccination studies

1.8.1 Molecular responses induced by timed vaccine perturbation

The earliest pioneering systems biology study using vaccination in 2007 by Fuller and colleagues profiled n=5 patients before and after a *Francisella tularensis* vaccine and defined modular patterns of genes with shared kinetics[182]. This identified “sustained up” pathways elevated past 2 weeks linked to proliferation, “down early” genes from

decreased frequencies of lymphocytes and “up early” mapping to myeloid cells and inflammation. These patterns broadly recur in the subsequent studies outlined below which further attempted to identify coherently induced modular patterns as well as those predictive of response quality. Vaccination studies measuring the transcriptome which focused on evaluating vaccine candidates[183,184] or defining minimal genes[185,186] as opposed to dissecting human variations are not reviewed in detail below. Antibody titers are often the response quality metric used, though specific response quality metrics unique to the biology of a certain vaccines are also defined. Even for highly efficacious vaccines like yellow fever, antibody and T cell responses can provide a measure of quantitative immunogenicity–despite the near universal protective nature of this vaccine, correlating molecular features with immunogenicity still provides a natural experimental model to understand quantitative variations in human immune systems.

1.8.2 Influenza

A study of influenza vaccination in 92 individuals by Bucasas and colleagues[187] was the first to adjust models for pre-existing influenza memory by defining the “titer response index” as the residual of a linear model adjusting for batch and baseline titer. Using cross validated mixed effects one way (time) ANOVA models without intrinsic covariate adjustment, the authors found the 24h post vaccination transcriptional response included IFN stimulated genes, IL6 pathway and JAK-STAT signaling genes. A gene signature predictive of titer response index included early IFN and antigen presentation genes and SPI1, a monocyte/macrophage lineage-determining transcription factor implicating monocyte frequency, cell intrinsic monocyte activation or both in the early response. Around the same time another group compared the trivalent inactivated vaccine (TIV) to the live attenuated influenza vaccine (LAIV)[188] which induces lower antibody titers. This study classified “high” response rate as individuals who met the World Health Organization criteria for response; 22/28 individuals receiving TIV were “high” responders. Many of these individuals would not be defined as high responders in other studies which focus on quantitative response level within the cohort, a difficulty of cross study comparison. Just 4/28 individuals receiving LAIV were “high” responders. Some ISGs induced by LAIV were hypothesized to reflect responses tailored live virus. Expression of interferon related genes 3 days post vaccination were correlated with high responders across both vaccines, and a supervised machine learning approach aimed at feature selection[189], identified a small number of genes

predictive of response, including day 7 signatures of activated B cells / plasmablast like TNFRSF17, but also day 3 expression of LST1, PYCARD, NLRP12, HSPA6, SPA6, LILRB2, LILRA1, LILRA3, and SIGLEC6. We investigated the enrichment of these signals against a compendia of databases[190] which suggested they are highly expressed in non-classical monocytes and are predicted to be regulated by SPI1. Early ISG / myeloid cell activation potentially linked to antibody response by Bucasas *et al.* [187] may thus persist to day 3. Intrinsic covariates were not adjusted for in this study.

The cohort profiled in chapters 2 and 3 were from an influenza vaccination study conducted at the NIH Clinical Center[39]. This report developed a metric to adjust for baseline antibody level while accounting for non-linearity between baseline titer and response. The adjusted maximum fold change (adjMFC) quantified relative response within bins defined by initial titer and represents response rate relative to individuals with similar baseline immunity. Coherent molecular responses were defined by adjusting transcriptome data for baseline expression, age, sex, batch, hidden batch effects[191] and ethnicity. Pathways enriched within coherently elevated genes on day 1 included PRRs, Fc-gamma Receptor-mediated phagocytosis, ISGs and NK cell signaling. Activated DC and monocyte subsets assessed by flow cytometry also increased coherently on day 1. Supervised diagonal linear discriminant analysis[192] defined predictive models and cross validation procedures defined robust correlations when predictive models were not achieved. ISG pathways on day 1 were robustly correlated with antibody titer maximum fold change, but not adjMFC. Predictive models could be built from both day 7 populations (B cell / antibody related genes) and more surprisingly, from baseline cell population frequencies described below (see section 1.8.8).

A unified meta-analysis demonstrated the robustness of the early inflammatory / interferon signatures using cohorts recruited at Yale over 5 seasons[162]. These authors developed a correction factor for baseline titers which built upon the titer response index[187] and adjMFC[39], but does not require pre-defined bins and uses an exponential function to account for nonlinearity. As described above (see section 1.6.1.3) early ISG and innate cell modules linked to response differed between old and young, consistent with prior studies[163]. The study identified the expected day 7 plasmablast signature but also interestingly day 28 expression of KLRB1 (CD161) positively correlated with response across 5 seasons in young adults but negatively associated with response in

older individuals. Whether this represented an NK cell or T cell frequency or state is unknown. Another integrative analysis[163] combined datasets with larger cohorts[134] and used an artificial neural network (ANN) classifier[193] to find predictive models irrespective of age. The ANN classifier selected a small number of predictive modules on 100 different trials, finding the expected day 7 B cell signatures identified in prior studies. Frequency of selected modules for each trial were not reported and the same module was selected at most ~20% of the time (Supplemental table 3 from Nakaya 2015), however clear patterns emerged with monocyte / receptors, DC activation and cell cycle processes with test / training accuracy ~ 70% were selected in nearly all trials. Defining the cellular origins of these signatures and understanding multicellular circuits and dynamics between baseline, early, and late expression profiles of influenza vaccination response are major goals of chapter 3.

1.8.3 Pneumococcus

Obermoser and colleagues[194] used expression of 62 gene modules they derived in a prior study[86] to compare early and late transcriptional responses elicited by the 23 valent unadjuvanted pneumococcal vaccine to TIV. The carbohydrate based pneumococcal vaccine did not induce any interferon response on day 1. Of 5 modules uniquely elevated by TIV, 4 captured ISG / IFN related processes and the 5th reflected myeloid cell activity or frequency recapitulating the major themes reviewed above. Further kinetic profiling showed TIV-induced ISGs were seen as early as 15h post vaccination. The authors sorted neutrophils, Monocytes, CD4 and CD8 T cells to attempt to localize the ISGs and found them preferentially up in neutrophils and monocytes, though notably they were also upregulated at lower magnitude in both T cell subsets at this early (24h) timepoint. Both vaccines induced modules related to inflammation and apoptosis on day 1. We further explore early apoptosis in chapter 3 in the context of adjuvant specific responses. Pneumococcal vaccine specific modules were related to inflammation, with driver genes including IL-1RN, TLR4, TLR5, TLR6, TLR8, CR1, MMP9 CD58, IL-8RA, CXCL1 and IL-1b. Both vaccines induced plasmablast frequency expansion correlating with B/plasma cell gene modules on day 7, with quantitatively higher day 7 fold change sustained out to day 10 in pneumococcal group. Carbohydrate based vaccination thus induced more inflammatory responses and larger total plasmablast expansion and persistence in blood compared to TIV.

1.8.4 Yellow Fever

The live yellow fever vaccine induces durable antibody and T cell response lasting decades[195]. It is possible that this replication competent vaccine may induce a second wave of responses after replication. A study from Emory university first reported the pan vaccine signature TNFRSF17 with other B cell related transcriptional responses on day 7 as correlates of later antibody response[144]. A cohort from Montreal measured different timepoints and found the plasmablast and B cell fold change appeared most induced day 14, as opposed to day 7[143]. In another cohort meta-analysis, we found indeed plasmablast kinetics were vaccine specific and were delayed following yellow fever vaccination with peak responses on day 14[196]. Early (day 3) response genes correlated with later T cell immunity in both studies appeared to reflect diverse metabolic processes and potentially innate cell activation of unknown origin (e.g. day 3 genes included activation (CD69) metabolism (ALDH3B1), complement (C1QB), and TLRs (TLR7, MYD88) though the cellular origins of these signals were unknown. The induction of a combined T helper 1/2 (TH1/TH2) profile is hallmark of YF vaccination, a process thought to relate to YF directly triggering multiple TLR pathways on DCs in mice[197] and in humans[198]. A more recent study of YF vaccination from China had increased temporal resolution at early timepoints including 4h and days 1, 2, 3, 5 and later timepoints[199]. This early time resolution identified genes linked to innate cell ontology and cytokines downregulated at 4 hours, potentially implicating early extravasation into tissues. Unsupervised co-expression networks[89] implicated both DCs and CD4 T cells in early responses, with innate cell and interferon response generally subsequently increasing and peaking around day 5, consistent with a second wave of responses to replication competent vaccines.

1.8.5 Malaria

The complex life cycle of *plasmodium falciparum* presents a challenge for current vaccines which lead to short lived immunity and fail to neutralize the liver stage, though promising candidates are on the horizon[200,201]. An early malaria challenge study[202] showed early induction of proteasome activation and IFNg 72h post vaccination could predict protection. A later challenge study compared 3 doses of RTS,S (RRR) with a 2 dose AS01 adjuvanted formulation after priming with a circumsporozoite expressing adenovector (ARR)[203]. Coherent transcriptional responses to both vaccines were similar and overlapped with the authors' prior studies of yellow fever[144], but differed in

signatures of protection. Similarly both vaccines had comparable efficacy, but different correlates of protection.

Antibody titer against circumsporozoite protein on the day of challenge corelated with protection for RRR. Surprisingly, RRR had an inverted pattern of antibody response signatures from the vaccines outlined above–plasmablast gene modules on day 1, and innate immunity modules related to antigen presentation, DC activation, and type I IFN on day 6, correlated with antibody response / challenge protection. These specific early plasmablast module genes were uncorrelated with plasmablast levels in serum, unlike the tight coupling seen after YF[144] and influenza[39] vaccination so may have derived from a different subset of activated B cells. Intriguingly in a study of VZV vaccination[204] (reviewed below, 1.8.6) a small subset of individuals had elevated baseline and an order of magnitude higher day 3 post vaccination induction of an inositol phosphate metabolism module. These individuals had peak plasmablast expression at day 1, and not day 7; in this context peak plasmablast response was linked to lower day 28 antibody levels, as well as lower activated CD4+ T cells on day 7, whereas the opposite was true in the inositol phosphate low group. In the context of studies showing early peak plasmablast activity on day 2 in elderly patients[164], together these results suggest early plasmablast activity but *not* peak activity, could reflect innate cell crosstalk with B cell subsets which mediates a positive feedback circuit sustaining IFN response genes later (d7) and associated with improved antibody response, as seen in RRR. However, early *peak* plasmablast responses may indicate a metabolic state coupled to hyporesponsiveness also linked to hyporesponse with age. Further studies may be warranted to investigate the effects of metabolic interventions targeting these pathways on immune responses in older individuals .

For RRS, frequency of CSP-specific polyfunctional CD4 T cells and not antibody level correlated with protection; modules related to antigen presentation, DC and monocyte activation and TLR signaling the first day after each dose were correlated with the level of CD4+ T-cell responses. This study also found a strong negative association between NK cell module activity on day 56 (28 days following boost); similar modules are coherently downregulated by both MF59 adjuvanted and unadjuvanted influenza vaccines but are not linked to antibody responses[205]. NK cell frequencies are variable and unstable at the population level[20,34], and as suggested by the authors, it is unknown if the negative association after malaria vaccination represents reduced NK cell

frequency due to enhanced NK activity in tissue[206], or negative regulation antibody responses[207] implicated by animal models.

1.8.6 VZV

The DNA virus Varicella Zoster (VZV) can reactivate to cause Herpes Zoster (shingles), a process thought to be linked to declining cellular immunity with age[208] Declining frequency of VZV specific T cells[208] and not antibody levels, are linked to reactivation. IFN producing VZV specific T cells are therefore utilized as a correlate of protection. The relevance of these circulating cells was demonstrated with a procedure involving subdermal injection of VZV antigens followed by punch biopsy[209]. VZV-specific CD4 T cells elevated in blood early after vaccination, but not tissue resident memory cells (TRM), were thought to be the origin of vaccine-responsive CD4 cells recovered from the biopsy based on their migratory gene expression signatures. The importance of circulating CD4 T cells in VZV responses was emphasized by a report which used supervised clustering of VZV specific IFN producing T cell kinetics pre and post vaccination over one month[210]. The authors found CD4 T cell attrition on day 28 from the peak at day 8 or 14 was critical for determining the increase in antigen specific T cells post vaccination. Monocyte activation genes day 1 post vaccination influenced expansion and subsequent persistence of effector T cells in a model of CD4 contraction kinetics. Sorted activated CD4 cells also demonstrated cell intrinsic cell cycle and proliferation transcriptome signatures which predicted T cell persistence in age adjusted linear models, indicating cell cycle checkpoints establish the extent of attrition of the predictive effector T cell populations.

Interestingly, a module positively linked to peak responses and negatively linked to contraction within CD4 T cells, was SREBF1 target genes (though this was not specifically commented on, see Fig 7A Qi *et al*.[210]) which are involved in fatty acid and carbohydrate metabolism. A later study of VZV vaccination used a multi-omic network integration model of bulk transcriptomics, metabolomics, cytokines and cell populations in response to VZV vaccination[204]. In this study, a multivariate model adjusting for sex, age, study site, and baseline VZV IgG, found SREBF1 activity positively associated with T cell responses at day 3 and antibody responses at day 7. Multiomic network analysis found the SREBF1 module was a major hub integrating intracellular programs and extracellular cytokine and metabolic signals across timepoints. These two studies demonstrate the benefit of directly profiling transcriptomes within sorted cells, since

SREBF1 signal could be pinpointed at least in part to the CD4 cells which were themselves the effectors of response quality.

1.8.7 Adjuvants

Adjuvants overcome hyporesponsiveness seen in the elderly and enhance responses through diverse mechanisms thought to occur at the axis of innate cell-lymphocyte information transfer[211]. MF59 (alum based) and AS03 (oil in emulsion, studied in chapter 3) are known to broaden the antibody repertoire[212] such that individuals make antibody to drift variants or even different strains of virus not included in the vaccine[213,214]. Based on these observations it was suggested these formulations may act directly on naive B cells[213] through an unknown mechanism. Previous molecular study of oil in emulsion adjuvants implicate their local action, which may involve ER stress[215] and IL2 producing CD4 T cells[216], early innate cell activation, and interferon responses[217,218]. Oil in emulsion adjuvants can also homogenize population variation in transcriptional responses[205] by inducing more uniformly strong responses; the degree to which different formulations have this property may be adjuvant dependent[219]. In a study of MF59 adjuvanted TIV, day 1 IFN and DC activation bulk transcriptome signatures were positively correlated with later antibody titers but interestingly were negatively corelated by day 3[205]. We further review and explore differences in adjuvanted vs unadjuvanted vaccines in chapter 3.

1.8.8 Baseline states predictive of functional response

Together, the studies above demonstrate how intrinsic variations and environmental factors can tune the molecular and cellular profile comprising the baseline immune system status. Pre-existing immunity measured through antibody titer forms another aspect of the baseline immune state emphasized above in the context of endpoint metric adjustment. Other latent factors likely contribute to baseline immune states unaccounted for by adjusting for baseline titer. An illustrative example from a study of an HIV vaccine candidate identified a distinct effect of pre-existing immunity on *molecular* responses[220]. Baseline transcriptomic modules linked to the outcome variable, cytotoxic CD8 responses, consisted of inflammatory cytokines, myeloid cell activation markers and ISGs. Surprisingly, baseline memory response to the adenovirus vector used in the vaccine was correlated with complete attenuation of early myeloid cell and interferon activation.

Despite these and other challenges accumulating evidence suggests a positive association between baseline interferon and potentially myeloid cell programs as well as B cell populations and later antibody responses to influenza vaccines. In the NIH vaccination study described above[39] transcriptional signals related to nucleotide metabolism, TREM1 signaling, Fc-gamma receptor signaling, PRR genes and ISGs were all robustly corelated with antibody response. In addition, several lymphocyte and innate subsets passed the strict criteria for being predictive of antibody response. Longitudinal analysis identified the predictive populations that were stable over time within individuals which potentially made them more indicative of a homeostatic immune "setpoint". Stable populations included CD161+ IL22 expressing T cells which were positively correlated with baseline PRRs TREM1 signaling ISGs gene. Population variation in basal ISGs with CD161+ cells producing TH-17 associated cytokines (IL-22) are interesting in light of the potential IFN related TH17 circuit implicated in the Immvar study of CD4 stimulation phenotypes[53]. The results implicate population variations in the interplay between interferon levels, lymphocytes with an inflammatory profile and innate cells which have been implicated in prior studies[221,222] and may induce mixed T helper lineages seen after yellow fever vaccination. These topics warrant further study.

Baseline CD38+ transitional B cell populations (independent of antigen specificity) were also predictive of antibody response[39], a finding later replicated in a yellow fever vaccination cohort[223]. In addition, Sobolev *et al.* found transitional B cell populations positively associated with adverse reactions to more potent adjuvanted vaccines[159]. Are these transitional B cell populations sentinels of a more globally responsive / reactive immune "setpoint" given their baseline link to these outcome phenotypes? We addressed this question in our work describing a shared immune setpoint between healthy individuals vaccination responses and SLE patients plasmablast associated flares[115]. In this work, we further investigated the biology of the predictive baseline B cell populations using baseline CITE-seq data from high and low influenza vaccine responders. First, by developing a bulk transcriptional surrogate of the predictive stable B cell populations using only <800 longitudinally stable genes, a ten gene signature (TGsig) predictive of antibody response across multiple influenza and yellow fever cohorts was defined. We then investigated whether this baseline signature linked to more robust B cell responses to vaccination could be linked to a molecular phenotype in B cell driven autoimmune disease. The level of TGsig at disease free timepoints was

correlated with elevated plasmablast activity during flares in a subset of SLE patients with B cell molecular profiles. We then utilized the denoising method described in chapter 2 on baseline CITE-seq data to define protein based immune subsets in high and low responders to pinpoint the TGsig genes. This analysis revealed a potential circuit driven by IFN hypothesized to derive from more activated pDCs driving paracrine activation of B cells through CD40-CD40L levels in T helper cells (Fig. 1.2).

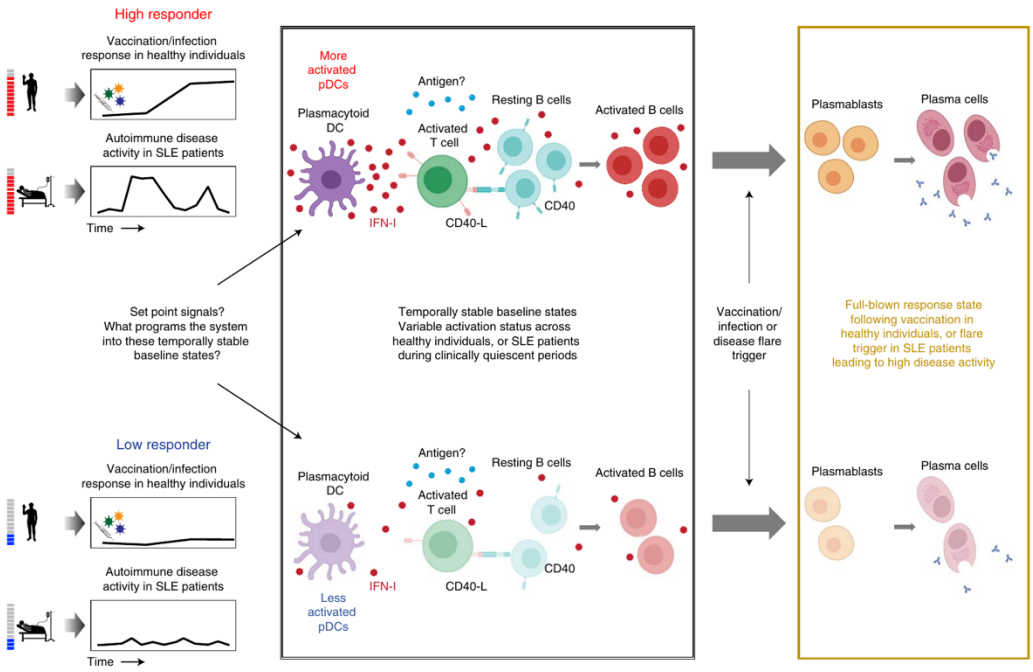


Figure 1.2 Inferred circuitries of a shared baseline immune setpoint.

Figure adapted from Fig.6b from Kotliarov *et al.* (2020). Model describing the molecular/cellular underpinnings and differences between high versus low responders. Activation of this entire circuit (including the components in the plasmablast/plasma cell box on the right) typically follows infection, vaccination, or occurs during autoimmune disease flares. Here we propose that the high responders tend to have more activated pDCs and thus more Type I IFNs and activated B and T cells at baseline, but only upon additional antigenic and/or inflammatory co-stimulation (and flare trigger in the case of SLE patients) does the system mount a full-blown plasmablast/plasma cell response cumulating in the generation of antibodies. Open questions include: 1) What sets the system into such temporally stable 'activated' states in pDCs, lymphocytes, and other myeloid cells? 2) What constrains the activated immune baselines from mounting full-blown plasmablast/plasma cell responses? 3) What is the antigen specificity repertoire of the activated lymphocytes at baseline?

How much might this setpoint generalize to other vaccines? While an early comparative study found baseline expression of monocyte modules negatively correlated response to influenza vaccination[163], the authors described challenges of batch effects but attempted

integration of data across all studies. Later, a meta-analysis strategy[224] used by the Human Immunology Project Consortium[161] used the adjMFC metric described above and implemented study level random effects on non-normalized data to borrow information across cohorts and improve generalzability[224]. This revealed baseline expression BCR signaling (M54), platelet activation (III) (M42) and inflammatory response (M33) were positively correlated with baseline adjusted antibody response in young individuals with FDR < 0.11, even after adjusting for cell frequencies in the validation cohort (the transcriptome and cell frequencies were measured in ref.[39] used for validation). These modules were not predictive of response in older individuals, indicating "inflammageing" signatures may be distinct from the inflammatory activation positively linked with response in the young. Subsequent analysis attempted to identify generalized baseline signatures across different vaccines. Class imbalance made universal prediction challenging, however unsupervised and supervised analysis honed in on innate cell inflammatory activation which we pinpointed to monocytes[225]. Related to baseline states predictive of antibody response and the divergence of signatures with age, a "bioage" signature reflecting finer shades of age related inflammation was negatively associated HBV vaccine response independent of sex and other clinical covariates[223]. The bioage score reflected decreased B cell signaling and increased inflammatory signaling (C1, MYD88, IRF7) at baseline. A later study[226] also found certain inflammatory regulators TNF and IL-9 at baseline negatively associated with HBV primary antibody responses, however baseline expression of CD14, LYN, IFITM3 (implicating activated monocytes) was associated with increased HBV antibody response. Furthermore post vaccination, these IFN induced states were elevated, suggesting monocyte ISGs may have been primed then elevated further by vaccination, possibly in younger individuals, with inflammation driven by TNF signaling reducing response in the elderly.

1.9 Concluding remarks

The studies reviewed above indicate molecular signatures related to intrinsic factors, coherent perturbations, and baseline states linked with response can be identified with top down systems biology approaches. The aim of these systems biology studies is to identify the precise molecular perturbations within cells that are linked to responses. A key limitation of the studies above is the limited resolution of bulk profiling which they all utilize. We extend multimodal single cell methods to a top down systems biology

study of vaccination to address this challenge. On the methodological front, a key advance includes identifying how noise in CITE-seq protein data is distinct from the type of noise observed in single cell mRNA data. We describe comprehensive deconvolution of these noise sources and our dedicated software method, “dsb”, in chapter 2. A second advance is development of a multiscale analysis framework for timed perturbation systems biology studies. We packaged multilevel modeling and various downstream functionality into another R package that can accommodate complex cohort single cell nested experiment designs[227]. This framework enables the identification of multicellular immune networks consisting of coupled transcriptional modules within cell types defined by (denoised) protein levels. In chapter 3 and 4 we use dsb and this software framework to help understand multicellular perturbation phenotypes induced by vaccination and immunotherapy, extend the concept of immune setpoints and understand their functional relevance. Unresolved questions which could not be answered without these approaches include 1. What are all unbiased states linked to perturbation responses within protein defined subsets? 2. What are the cellular origins of past signatures bulk expression at the gene level module level and single cell level? 3.What are the post vaccination kinetics of baseline states linked to the response and are those same cells perturbed by vaccination? Finally, by extending this framework to the setting of cancer immunotherapy, we ask whether setpoints and response phenotypes can be identified and linked to other clinical outcomes, such as adverse events.

2 NORMALIZING AND DENOISING PROTEIN EXPRESSION DATA FROM DROPLET-BASED SINGLE CELL PROFILING

This work was published in Nature Communications as:

Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. Nat. Commun. 13, 2099 (2022).

https://www.nature.com/articles/s41467-022-29356-8

DOI: 10.1038/s41467-022-29356-8

The figures referenced as "Supplementary Fig" refer to the figures shown in the Appendix. The appendix of this thesis is the 'supplementary note" for the article above as referenced in the chapter below as "appendix". The appendix can also be found here.

2.1 Abstract

Multimodal single-cell profiling methods that measure protein expression with oligo-conjugated antibodies hold promise for comprehensive dissection of cellular heterogeneity, yet the resulting protein counts have substantial technical noise that can mask biological variations. Here we integrate experiments and computational analyses to reveal two major noise sources and develop a method called "dsb" (denoised and scaled by background) to normalize and denoise droplet-based protein expression data. We discover that protein-specific noise originates from unbound antibodies encapsulated during droplet generation; this noise can thus be accurately estimated and

corrected by utilizing protein levels in empty droplets. We also find that isotype control antibodies and the background protein population average in each cell exhibit significant correlations across single cells, we thus use their shared variance to correct for cell-to-cell technical noise in each cell. We validate these findings by analyzing the performance of dsb in eight independent datasets spanning multiple technologies, including CITE-seq, ASAP-seq, and TEA-seq. Compared to existing normalization methods, our approach improves downstream analyses by better unmasking biologically meaningful cell populations. Our method is available as an open-source R package that interfaces easily with existing single cell software platforms such as Seurat, Bioconductor, and Scanpy and can be accessed at "dsb [https://cran.r-project.org/package=dsb]".

2.2 Introduction

Recent developments in multimodal single cell analysis involve using DNA barcoded antibodies to simultaneously profile surface proteins together with the transcriptome (e.g., CITE-seq) and/or chromatin accessibility (e.g., ASAP-seq) in single cells[44–47]. This greatly enhances our ability to discover, define, and interpret cell types and states, particularly those comprising the immune system given extensive existing knowledge connecting surface protein profiles to immune cell subsets and functions[9]. Droplet-based sequencing of single cells stained with DNA-barcoded antibodies provides a readout of protein levels in the form of antibody-derived tag (ADT) counts for each protein. This “cytometry via sequencing” approach bypasses spectral interference inherent in fluorescence-based cytometry methods, thus enabling simultaneous profiling of hundreds of proteins in single cells. While low-level normalization and modeling approaches for single cell RNA-seq data have received considerable attention[228–234], those for protein/ADT are in their infancy and more importantly, the extent and sources of noise have not been quantitatively analyzed despite the substantial levels of apparent noise reported in raw protein counts[44].

Stochastic processes during single cell mRNA capture and sequencing contribute to sampling noise[235,236] and other technical variations leading to reduced UMI counts, including zero counts for genes despite actual mRNA expression in a given cell. Such noise can be modeled with statistical distributions[237–239] or normalized, for example, by standardizing the total number of mRNA reads between cells commonly performed via scaling factors computed from each cell’s total mRNA “library size”[240] (defined herein

as the total UMI count for a given assay/data modality in each cell). However, these methods are not appropriate for surface protein count data for several reasons. First, a major noise component of ADT data appears to be added background noise because cells tend to have positive counts for multiple classes of proteins that are reported to be mutually exclusively expressed in distinct cell subsets. For example, compared to sparse mRNA counts, only two 0 values are present across more than 11,000 cord blood cells stained with 13 surface proteins in the original report of the CITE-seq method[44]. Second, current methods/experiments still measure only a small fraction of unique proteins with a wide range of antigen density on different cell types, resulting in individual protein counts in single cells spanning ~2-3 orders of magnitude (e.g., less than 10 to more than 1000); differences in total protein counts between cells therefore depend on the specific antibody panel used. Finally, the total protein counts detected on a given cell may reflect both technical but also biological variations such as cell size across cells and cell types, especially given the dependence of the total ADT counts on the specific antibody panel used.

The original developers of CITE-seq normalized ADT data by using a centered log ratio transformation (CLR). The resulting values can be interpreted as either a natural log ratio of the count for a given protein relative to the other proteins in the cell (CLR "across proteins", as implemented in the original report of CITE-seq[44]) or relative to other cells (CLR "across cells", a modification used in later work by the authors[98], which renders CLR less dependent on the composition of the antibody panel). The CLR transformation helps to better separate cell populations, but it does not directly estimate and correct for specific sources of technical noise including the apparent background noise mentioned earlier. The authors accounted for protein-specific noise in human cells by spiking in mouse cells to set a per-protein cutoff for determining whether a CLR transformed (across-protein) expression value was above that in mouse cells[44]. This approach appears not adopted beyond its original use, likely because it entails more complex experiments and analyses. More recent reports applied other approaches, for example, fitting models to estimate background and foreground distributions for each protein without using spike-in control cells[241,242], or using isotype antibody controls to estimate background[47,243]. It is unclear the extent to which these approaches remove noise versus biologically relevant signals since the noise sources remain unidentified; some proteins also have multimodal distributions across cells, while isotype controls are not typically used in flow cytometry for quantitative thresholding[244] since their level can

reflect both biological and technical variations. Thus, determining the major sources of noise and developing dedicated methods to account for them are major unmet needs given the swift adoption and proliferation of multimodal single cell profiling methods involving the measurement of protein expression with DNA barcoded antibodies.

Here we perform experiments and computational analyses to reveal two major components of protein expression noise in droplet-based single cell experiments: 1) protein-specific noise originating from ambient, unbound antibody encapsulated in droplets that can be accurately estimated via the level of "ambient" ADT counts in empty droplets, and 2) droplet/cell specific noise revealed via the shared variance component associated with isotype antibody controls and background protein counts in each cell. We develop an R software package, "dsb" (**d**enoised and **s**caled by **b**ackground), the first dedicated low-level normalization method developed for protein ADT data, to correct for both of these noise sources without experimental modifications. Our application of this approach to our own and several external data sets spanning multiple technologies and assay types demonstrates the generalizability of dsb to enhance downstream analysis, including manual and unsupervised protein-based and multimodal (joint protein-mRNA) identification of cell populations and states.

2.3 Results

2.4 Analysis of unstained cells reveals ambient antibody capture as a major source of protein-specific noise

To assess protein count noise, we first utilized our previously reported dataset measuring more than 50,000 Peripheral Mononuclear Cells (PBMCs) from 20 healthy human donors[115] stained with an 87 CITE-seq antibody panel (including four isotype controls; Totalseq-A reagents, Biolegend). Consistent with the original CITE-seq report[44], we noticed non-zero counts for most proteins in each cell, resulting in positive counts even of markers not expected to be expressed in certain cell types. We also noticed non-zero, "ambient" protein counts in tens of thousands of empty droplets containing capture beads without cells, which emerge naturally due to Poisson distributed cell loading, reminiscent of cell-free RNA observed in droplet-based single cell RNAseq[245–247]. We reasoned that background noise in CITE-seq data may partly reflect such unbound, ambient antibodies captured in droplets. To assess whether counts

in empty droplets indeed reflect the ambient component in cell-containing droplets, we compared background protein levels in cell-free droplets with droplets capturing unstained control cells spiked into the cell mixture after cell staining and washing but prior to droplet generation (Fig. 2.1a). We found positive protein counts even for unstained control cells, and that the average log transformed level per protein in empty droplets and unstained control cells were highly correlated (Fig. 2.1b). A similarly strong correlation was observed between the average protein counts in subpopulations of stained cells "negative" for a given protein and those in empty droplets (Supplementary Figs. 1a-c; negative cells correspond to those in the fraction with lower expression of the protein–see Methods), further suggesting that the noise component correlated across cells is dominated by ambient antibody capture. Thus, protein counts in empty droplets, which are available in all single cell droplet experiments, provide a direct estimate of the ambient background due to free antibody capture for each protein. Consistent with our findings on ambient antibody capture as the major source of background noise in CITE-seq data, a recent study reporting CITE-seq antibody titration experiments across a wide concentration range demonstrated that background noise increased with the antibody staining concentration, with some antibodies at or above 2.5μg/mL having even more cumulative UMIs in the empty droplets compared to cells[248]. Our observation thus motivated the first step of our method to remove protein-specific technical noise: transforming counts of each protein in cell-containing droplets by subtracting the mean and dividing by the standard deviation of that same protein across empty droplets (see Methods). The resulting transformed protein expression values for each cell reflect the number of standard deviations above the expected ambient capture noise, thus centering the negative cell population for each protein around zero to help improve interpretability of the resulting protein expression values (Fig 2.1c).

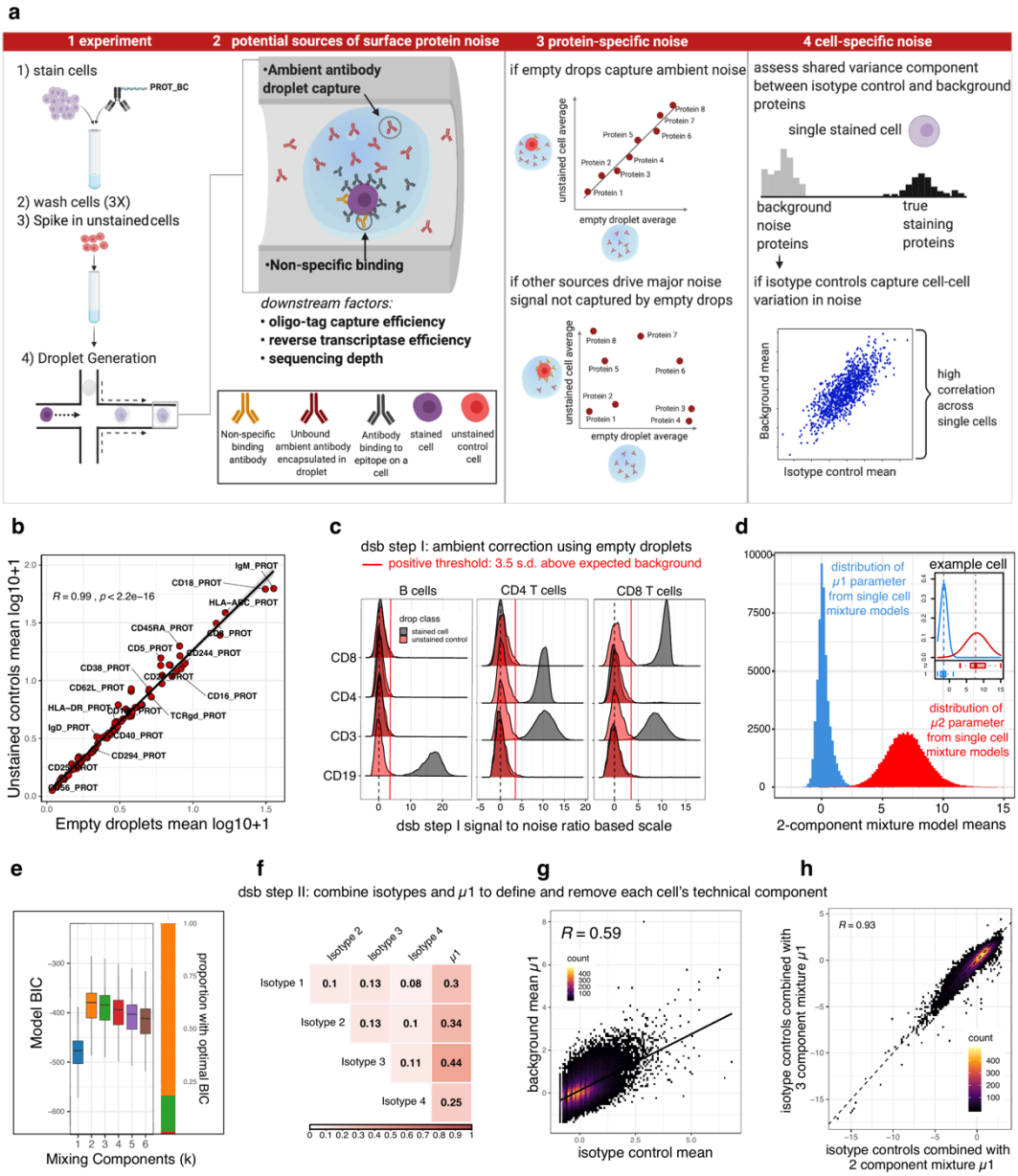


Figure 2.1 Antibody derived protein UMI count data noise source assessment.

a. 1-2: Experimental setup and potential noise sources in CITE-seq data. 3: protein-specific noise: if ambient antibody encapsulated in droplets constitutes a major source of protein-specific noise, values should be highly correlated with those in unstained control cells (top); if control cells contain information on noise not captured by empty drops, the correlation should be weak. 4: Cell-specific noise evaluated through the correlation between the background protein population mean and isotype controls across single cells. Created with BioRender.com. **b.** Average protein log10(count+1) of unstained control cells spiked into the stained cell pool prior to droplet generation (y-axis) versus that of droplets without a cell (x-axis). Pearson correlation coefficient and p value (two sided) are shown. **c.** Density histograms of protein expression of lineage-defining proteins within major subsets in stained cells (black) and unstained controls (red) normalized together using dsb step I (ambient correction and rescaling based on levels in empty droplets). **d.** A two-component Gaussian mixture model was fitted to

the protein counts within each single cell; the distributions of the component means from all single cell fits (blue=”negative” population; red=”positive” population) are shown, protein distributions from a randomly-selected cell shown in the inset. **e.** Comparison of Gaussian mixture models fit with between k = 1 and k = 6 subpopulations to dsb normalized protein values for n = 28,229 cells from batch 1 after dsb step I (ambient correction) but prior to step II, vs the model fit Bayesian Information Criteria (BIC, using mclust R package definition of BIC where larger values correspond to a better fit) from the resulting 169,374 models. Boxplots show the median with hinges at the 25th and 75th percentile, whiskers extend plus or minus 1.5 times the inter quartile range. k=2 component Gaussian mixtures have the best fit in more than 80% of cells (orange, right inset bar plot). **f.** Pearson correlation coefficients among isotype controls and background component mean inferred by Gaussian mixture model (μ1 fitted per cell as in d); all corresponding p values (two sided) are less than 2e-16. **g.** Scatter density plot between μ1, the mean of each cell’s negative subpopulation from the per-cell Gaussian mixture model (blue in Fig. 2.1c) versus the mean of the four isotype controls across single cells. Pearson correlation coefficient is shown (two-sided p value < 2e-16). **h.** The distribution of the dsb technical component as calculated using a 2 component (x-axis) vs. 3 component (y-axis) mixture model to define the μ1 parameter, Pearson correlation coefficient, p value (two sided) < 2e-16.

2.5 Shared variance between isotype controls and background protein counts in single cells provide cell-intrinsic normalization factors

In addition to ambient noise correlated across single cells as captured by average readouts from empty droplets, cell/droplet-intrinsic technical factors including but not limited to oligo tag capture, cell lysis, reverse transcriptase efficiency, sequencing depth and non-specific antibody binding, can contribute to cell-to-cell variations in protein counts that should ideally be normalized across single cells. Given that the differences in total protein UMI counts between individual cells could reflect biologically relevant variations such as those due to the physical size of naïve vs. activated lymphocytes, library size normalization (dividing each cell by the total library size) could remove biological rather than technical cell to cell variations. In addition, since current CITE-seq antibody panels are a small subset of total surface proteins, the assumption that total UMI counts should be similar among cells may not be valid. Here we integrated two types of independently derived measures to reveal a more conservative (i.e., avoiding over-correction and removal of biological information), robust estimate of the factor associated with cell-intrinsic technical noise (Fig. 2.1a).

First, the four isotype control antibodies with non-human antigen specificities in our panel could in principle help capture contributions from non-specific binding and other technical factors discussed above. The counts of the isotype controls were only weakly (but significantly) correlated with each other across cells (Fig. 2.1f), and interestingly, the correlation between the mean of four isotype controls and the protein library size (which has both biological and technical components) across single cells was even higher (Pearson correlation 0.45) than that between the protein library size and the individual isotype controls (average Pearson correlation 0.25). This suggests that while each isotype control may be individually noisy, and their levels may still partially reflect biological contributions, collectively their shared component of variation (i.e., as reflected by their average) may better capture technical noise in the experiment. Second, to further boost the robustness of estimating cell-intrinsic technical noise, particularly given that the number of isotype controls available in practice can be limited, we sought an additional estimate of droplet-intrinsic technical variation. Since each cell in a sample of multiple distinct cell types (e.g., PBMCs) is expected to express only a subset of protein markers in staining panels, we reasoned that the distribution of each cell's non-staining proteins (e.g., those specifically expressed in other cell types/lineages) could be differentiated from the cell's "positively expressed" proteins by fitting a 2-component mixture model to each cell. If so, the average counts in the population of non-staining/negative proteins could reflect and therefore serve as another readout of the cell's technical component that could then be integrated with the cell-intrinsic noise captured by isotype controls. To assess this hypothesis, we applied a Gaussian mixture model with two (k=2) subpopulations to fit the protein counts within each single cell after correcting for the protein-specific ambient noise we identified above (see below and Methods; Fig. 2.1d). We found clear separation between the background (with mean=$\mu 1$) and positive (mean=$\mu 2$) protein population with substantial cell-to-cell heterogeneity of subpopulation means (Fig. 2.1d). We next assessed the robustness of using a 2-component mixture to model the protein counts of individual cells by comparing k=1 to 6 component models assessed using the Bayesian Information Criterion (BIC). While two-component models had the best fit in a majority (81%) of cells, indicating a bimodal protein distribution within single cells, k=3 models had the best fit in nearly all remaining cells (Fig. 2.1e, Supplementary Figs. 2a-b; see also Supplementary Note). The BIC for these cells were very similar to the corresponding k=2 models (Supplementary Fig. 2c), indicating that the 2-component fits were

identifying very similar positive and negative populations. Importantly, for the minority of cells with optimal k=3 or 4 models, the resulting mean of the lowest expression population (μ1 estimate) was highly concordant when the same cells were fit with a k=2 model (Supplementary Figs. 2d-f). These data suggest that a 2-component Gaussian mixture fit of the protein population within single cells can robustly delineate the negative background protein count population for most cells.

Together μ1 and the isotype controls provide estimates of technical noise within each single cell. However, each variable may be individually noisy; we thus assessed information sharing among these variables. The correlations between μ1 and each individual isotype control (average correlation r = 0.33) or the average of all four isotype controls (r=0.59) were higher than those between the isotype control themselves (average correlation r = 0.11), suggesting that the shared variation (i.e., average) between the independently inferred μ1 and isotype controls captured unobserved, latent factors contributing to technical noise (Figs. 2.1f-g). We thus reasoned that the first principal component score (λ) capturing the shared variation of μ1 and the isotype controls across single cells would be a robust measure of technical noise intrinsic to individual cells. λ was associated with the protein library size across single cells within cell clusters (Supplementary Fig. 3a clusters defined after dsb steps I and II, see Methods), supporting the notion that λ captures the technical component of the protein library size. Furthermore, consistent with the observation above regarding the Gaussian mixture model fit, λ was highly concordant regardless of whether the background (μ1 estimate) was defined using a k=2 or k=3-component Gaussian mixture (Fig. 2.1h).

Given the information sharing between μ1 and isotype controls, we recommend the inclusion of multiple isotype controls in CITE-seq experiments to serve as anchors for robust inference of technical normalization factors (see Supplementary Note). Together, our data indicate that while the signal from individual measures such as isotype controls can be noisy and may reflect multiple yet often unknown sources of variation, their correlated component of variation can serve as a robust normalization factor for surface protein expression in single cells. Thus, in a second, optional but recommended step, our method computes λ for each cell as its “technical component”, which is then regressed out of the ambient noise corrected protein values (Fig. 2.1c) generated by step 1 above (see Methods). The underlying modeling assumptions of the dsb technical component also held well in seven independent datasets generated via different assay

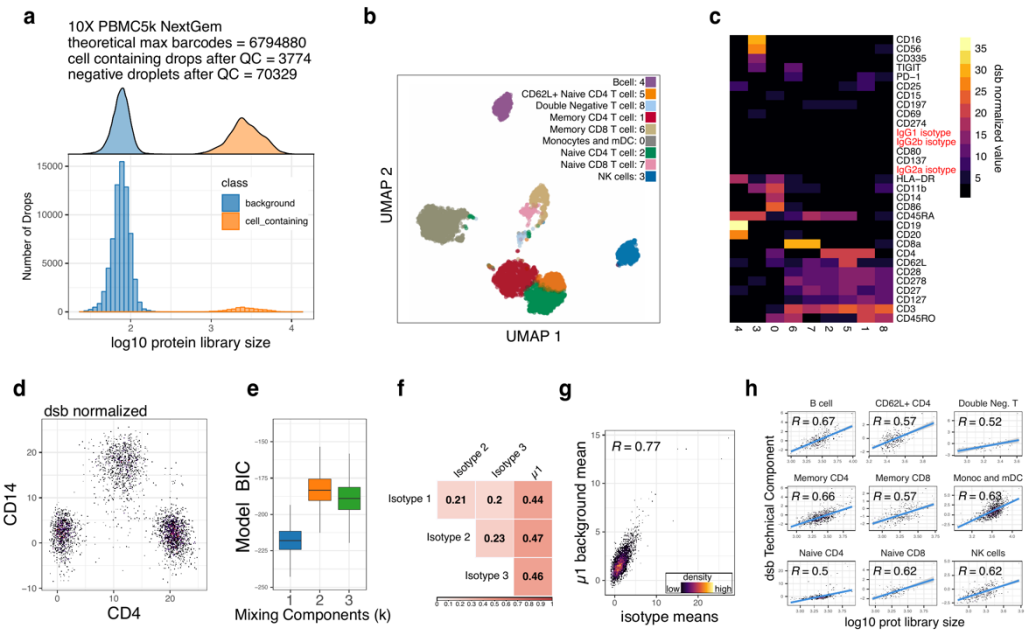
platforms and protein panels of diverse sizes (from 17 to more than 200 proteins; see below).

2.6 Comparison with other transformations and assessing dsb in independent datasets generated by different technology platforms

The unstained spike-in cells above should reflect the level of protein specific, "ground-truth" noise, we thus used these cells to visually compare dsb with other normalization transformations (Supplementary Fig 3e; see Methods). Unstained cells normalized using dsb centered around zero, while CLR or log transformation placed these cells at arbitrary locations. For example, CD4 has a trimodal distribution due to absence of expression in populations such as B lymphocytes, low expression in CD14+ monocytes and high expression in helper T cells; dsb normalized values centered the background population together with unstained control cells at zero and delineated low-level CD4 staining on monocytes. In contrast, these monocytes are closer to and partially overlapped with the unstained population when CLR or log normalization were used (Supplementary Fig 3e). We further compared dsb to CLR (the version that normalizes across cells) since CLR is the most commonly applied transformation for ADT data to date and normalization across cells should depend less on the protein staining panel than CLR across proteins. Using k-medoids clustering of single cells based on protein expression data only, the Gap-Statistic[249], which reflects improvement in within-cluster coherence relative to that expected of random data drawn from a reference distribution, was consistently higher using dsb than CLR across different values of k. However, the trend as a function of k was similar between dsb and CLR, suggesting that the improvement could be partly due to scaling differences between these two transformations (Supplementary Fig. 3f). Finally, differential expression analysis comparing major immune cell populations with the rest of the cells revealed that key lineage and cell-type specific proteins (e.g., CD56 on NK cells) tended to have larger fold changes when using dsb normalized protein values compared to CLR (Supplementary Fig. 3g).

We next tested the general applicability of dsb by using several independent, publicly available CITE-seq datasets. We first assessed whether the modeling assumptions developed using our own CITE-seq data would generalize to four other CITE-seq datasets that profiled ~5,000 to 10,000 cells using 14–29 surface phenotyping proteins and 3 isotype controls, and were generated using different versions of the 10X

Genomics droplet profiling kit than the one we used. Similar to our dataset, we detected a large number of empty droplets containing antibody reads (>50,000) inferred by the EmptyDrops[245] algorithm used in the Cell Ranger barcode rank algorithm; the number of cell-containing droplets estimated by Cell Ranger and further filtered by quality control metrics (3,000-8,000 droplets) was also consistent with the number of loaded cells (Fig. 2.2a, Supplementary Figs. 4a,h,o). Thus, protein-specific ambient noise can be estimated as in our data set using these empty droplets. Applying dsb without any modification resulted in biologically interpretable protein-based clusters (Figs. 2b,c, Supplementary Figs. 4e-f,l-m,s-t) and canonical immune cell populations could be clearly delineated by conventional biaxial plots (Fig. 2.2d, Supplementary Figs. 4g,n,u). Importantly, the model fitting behavior and correlations among isotype controls and background counts observed in our dataset were similarly observed in these independent datasets, including: 1) The k=2 component Gaussian mixture model had the best fit according to BIC in most single cells (Fig. 2.2e, 89% average across 4 CITE-seq datasets); 2) the estimated $\mu 1$ (mean of background protein counts) for each cell correlated significantly with the mean of isotype controls across single cells and was higher than the correlation with individual isotype controls (Figs. 2.2f,g, Supplementary Figs. 4b-c,i-j,p-q); 3) the inferred technical component using isotype controls and $\mu 1$ was correlated with the protein library size (Fig. 2.2h, Supplementary Figs. 4d,k,r); finally, 4) even on the smallest panel (14 phenotyping antibodies, 3 isotype controls) the per cell technical component λ was highly concordant regardless of whether the background ($\mu 1$ estimate) was defined using a k=2 or k=3-component Gaussian mixture (Supplementary Fig. 2.2g).



i

Dataset	Source	n phenotyping proteins	n isotype controls	n cells	n background drops used by dsb	Cor 1	Cor 2
Healthy Donor pre-vaccination (PBMC)	PMID: 32094927 / this study	83	4	53201	18529	0.61	0.59
ASAP-seq (Bone Marrow Aspirate)	PMID: 34083792	238	4	10927	55275	0.62	0.58
TEA-seq (PMN-depleted PBMC)	PMID: 33835024	45	1	28961	240352	0.41	0.34
10X PBMC 10K V3 (PBMC)	10X Genomics	14	3	7087	69481	0.57	0.77
10X PBMC 5` 5K V2 (PBMC)	10X Genomics	14	3	7089	55541	0.62	0.86
10X PBMC5k NextGem (PBMC)	10X Genomics	28	3	3774	70329	0.64	0.77
10X PBMC5k V3 (PBMC)	10X Genomics	28	3	3481	67204	0.7	0.74
Leukemia derived cell line mixture	Mission Bio	9	1	1049	16455	NA	NA

Cor 1 = dsb Technical component vs log protein library size
Cor 2 = cell Isotype control mean vs mixture model background mean

Figure 2.2 Assessment of dsb model assumptions and performance of dsb normalization on external datasets.

Panels **a-h**: application of dsb to a publicly available dataset generated using 10X genomics "NextGem" chemistry measuring 29 proteins across ~5K cells. **a.** The protein library size distribution of empty and cell-containing droplets used for dsb normalization. **b.** UMAP of single cells based on dsb normalized protein values with colors representing clusters obtained from clustering cells on dsb normalized protein values. **c**. Heatmap of the average of dsb normalized values per protein-based cluster shown in (b). **d.** The distribution of CD14 and CD4 dsb normalized values. **e.** As in Fig. 2.1e, Gaussian mixture model parameters fit to the dsb normalized values of each single cell after step I (ambient noise/background droplet based correction). The Bayesian Information Criterion (BIC) of the model vs. number of components in the model fit for each cell (n=3774 cells). Boxplots show the median with hinges at the 25th and 75th percentile and whiskers extending plus or minus 1.5 times the inter quartile range. **f.** As in Fig. 2.1f, Pearson correlation coefficient matrix of variables used to define each cell's technical component; each isotype control and μ1, the Gaussian mixture model background mean across proteins for each cell. **g.** As in Fig. 2.1g, Pearson correlation coefficient between the inferred cell-specific background mean μ1 from the Gaussian mixture model vs. the mean of isotype controls in each cell. **h.** The relationship between each cell's technical component and the cell's protein library size (Pearson correlation coefficient shown as in Supplementary Fig 3a with 95% confidence interval in grey). **i.** Summary statistics for the eight independent datasets assessed in this study; Cor 1 and 2 correspond to the Pearson correlation coefficient for assessing the relationships between variables shown in (h) and (g) across cells for each dataset.

We next tested the applicability of dsb to several new types of multimodal single cell data generated by technologies that measure surface protein expression in droplet captured single cells using oligo-barcoded antibodies including 1) "proteogenomic" data (protein + DNA mutation assays from Mission Bio; 9 proteins plus an isotype control), 2) ATAC-seq with Select Antigen Profiling (ASAP-seq: protein and chromatin accessibility; 238 proteins plus isotype controls), and 3) Transcription, Epitopes, and Accessibility (TEA-seq: protein + chromatin accessibility and transcriptome assessment; 45 proteins plus one isotype control). All datasets had ADT reads in a large number of empty droplets (Supplementary Figs. 5a,d,e). Our method was compatible with the proteogenomic dataset, helping to identify markers for each cell cluster after correcting for protein-specific background levels estimated from >16,000 empty droplets (Supplementary Figs. 5a-c). In the ASAP-seq dataset that measured multiple isotype controls, $\mu 1$ again correlated significantly with the mean of isotype controls across single cells and this correlation was higher than that among the individual isotype controls (Supplementary Figs. 5f,g), and the inferred per-cell dsb technical component was correlated with the library size as observed above (Supplementary Fig. 5h). In TEA-seq and ASAP-seq data, the negative staining cells could often be identified by applying the same 3.5 threshold that we applied in our and other data sets (Supplementary Figs. 5i-k and see below). The compatibility and utility of dsb with large protein panels such as in the ASAP-seq dataset is consistent with our recent CITE-seq analysis of Covid-19 patients using a similarly large panel where dsb helped enable accurate cell population identification by both automated clustering and manual gating[250]. A summary of results from these datasets is shown in Fig. 2.2i.

2.7 Case study I: dsb improves interpretation of protein-based and joint protein-mRNA clustering results

We next further investigated the ways in which normalization with dsb could help improve cell type identification. By design, dsb zero-centers the background population for each protein and provides normalized expression interpretable as signal above expected background noise. These features are thus particularly helpful in manual

gating across cell lineages (Supplementary Fig. 6a) and can improve the annotation of cell types derived from unbiased clustering. In contrast, distinguishing true biological expression from noise within individual cell clusters can be challenging when using transformations such as the CLR, partly because CLR protein values lie on a non-zero-centered scale (each protein also has a distinct noise floor); therefore, cells can appear to express markers known to be specific for other cell lineages. For example, in cluster 4 from our PBMC data (framed cluster in Fig. 2.3a), proteins such as IgA/IgM and CD57 could be mis-interpreted as showing signal above noise (Fig. 2.3b). In contrast, dsb normalized values for IgA, IgM, and CD57 are zero-centered (Fig. 2.3b), indicating that the level of these proteins in this cluster was statistically similar to the level in empty droplets and were therefore not expressed (Figs. 2.3c,d–red proteins). In contrast, CD16, CD244, and CD56 had dsb values above 8 (i.e., greater than 8 standard deviations above the mean in empty droplets, +/- the correction from regressing out the technical component), suggesting these were CD57 negative CD16++CD56+ NK cells, which are not known to express B-cell markers such as IgM or IgA. In general, cell clusters identified using dsb normalized protein values had cell type-defining proteins detected above the same threshold (3.5) applied within each cell cluster (Fig. 2.3e, Supplementary Figs. 6b–c).

We also assessed compatibility of dsb with an unsupervised joint mRNA-protein clustering algorithm that constructs a weighted nearest-neighbor (WNN) joint embedding of CITE-seq mRNA and protein data[251] (Fig. 2.3f). We ran WNN clustering using the same processed mRNA data together with ADT data normalized by either dsb or CLR (across cells). The clustering results were similar, suggesting dsb and CLR led to broadly concordant results. However, closer examination of individual clusters revealed that dsb could lead to more interpretable results. Notably, CD14 positive cells (presumably monocytes) were distributed across multiple dsb-derived clusters, including cluster 3 characterized by elevated CD86 (Fig. 2.3g). In contrast, the CLR value of these same cells was relatively low for CD86 but high for other markers (e.g., CD8 and IgM) that should not be expressed by monocytes (Fig. 2.3h). Furthermore, median CLR values in these cells (but not dsb – Fig. 2.3g) were correlated with the 98th percentile of expression in empty droplets across proteins ($R = 0.67$, $p = 3.1e\text{-}10$; Fig. 3h), suggesting that protein-specific ambient noise contributed substantially to the CLR values; this noise source was successfully accounted for by dsb via the use of empty droplets. Finally, relative to the rest of the cells, differentially expressed transcripts in

dsb-derived cluster 3 include inflammatory and activation genes (Fig. 2.3i), consistent with the CD86-high phenotype revealed by dsb.

Figure 3

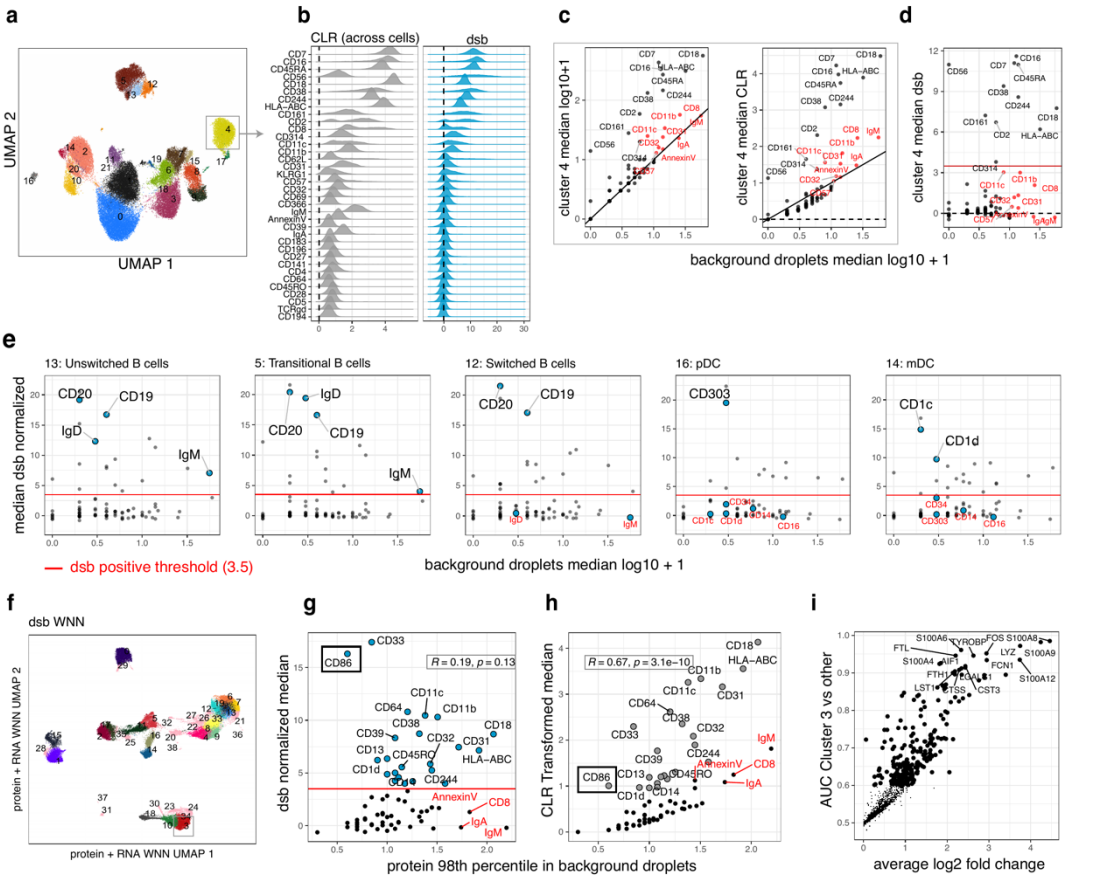


Figure 2.3 Case study I: dsb improves interpretation of cell clusters derived from protein-based and joint mRNA-protein clustering.

a. UMAP plot of single cells labeled by cluster number (clustering was performed using dsb normalized protein values). **b**. The distribution of protein expression of cluster 4 (highlighted with a grey box in (a)) using CLR (across cells) or dsb for normalization. **c**. Median log + 1 protein levels (left) and CLR transformed across cells (as in (b), right) in cells from cluster 4 versus the level in empty droplets; proteins highlighted in red are comparable in expression to "positive" proteins after log transformation (left) and CLR transformation across cells (right) but are similar to background levels in empty droplets (identity line y = x shown in black). All proteins with median log10 expression greater than 1 but less than 3.5 after dsb normalization are labeled with the protein name. **d.** Similar to (c), but the y axis shows the median dsb normalized values; proteins in red (those near the diagonal in (c)) are now residing below our uniformly applied dsb positivity threshold of 3.5, reflective their proximity with mean counts in empty droplets; proteins above the red line have median dsb normalized expression within the highlighted cluster 4 (see (a) and (b)) above 3.5, i.e., 3.5 standard deviations above ambient noise, +/- adjustment for the cell intrinsic technical component. **e.** The dsb normalized value vs. the median value in empty droplets of proteins within a subset of protein-defined clusters. A subset of proteins informative for cluster identification from B cell and dendritic cell subsets with a dsb value above 3.5 (red line) are annotated with

the protein name within each panel and are labeled in red when below 3.5 within each subset. Proteins labeled for B cell subsets (C13: Unswitched B cells, C5 Transitional B cells, C12 Switched B cells) include B cell proteins CD20, CD19, IgD, and IgM, proteins labeled for the dendritic cell subsets (C16: pDC, C14: mDC) include innate cell markers CD1c, CD1d, CD34, CD14, CD16, and CD303. **f.** UMAP plot of the same cells shown in (a) but the UMAP embeddings and clusters here were derived using Seurat's weighted nearest neighbor (WNN) mRNA-protein multimodal algorithm applied to dsb normalized values. **g.** Similar to (d) but derived using cells from WNN cluster 3; Pearson correlation coefficient and p value (two sided) are shown between median dsb normalized values and the 98th percentile expression value (log10) of the same protein in empty droplets. **h.** Similar to (g) but for CLR normalized values. **i.** Differentially expressed genes (ROC test; see Methods) for cell in cluster 3 vs other clusters.

2.8 Case study II: dsb unmasks MAIT cell population in tri-modal TEA-seq data

As a second example, we further analyzed trimodal transcriptome, protein, and chromatin accessibility (TEA-seq) data[252]. Visual inspection suggested improvement in biaxial plots after dsb normalization as the same interpretable threshold of 3.5 applied to all datasets in this study delineated two cell populations based on CD4 and CD14 (Fig. 2.4a) compared to normalization with protein library size (as implemented in the original TEA-seq study), and CLR (Supplementary Fig. 9a-b). To assess unsupervised multimodal clustering, we carried out the same comparison of CLR and dsb normalization using WNN clustering (combining mRNA and protein) as above but on TEA-seq data. Similar to above, the clustering results overlapped significantly (Chi-squared test, $p < 2e\text{-}16$ Supplementary Fig. 9 c-d). However, we noticed phenotypic marker differences within a specific T cell cluster that could substantially change the biological interpretation of the resulting cell population. During thymic development, human T cells rearrange variable, diversity and joining (VDJ) genes at the T Cell Receptor (TCR) locus. The resulting TCR gene rearrangements are distinct to functional categories of T cells with known specialized functions. This TEA-seq data included antibodies specific for alpha-beta (TCR a/b - conventional helper and cytotoxic T cells), gamma-delta (TCR g/d gamma-delta T cells), and Va7.2 (specific for mucosal associated invariant T (MAIT) cells). The MAIT TCR Va7.2 median dsb values were high (~15 standard deviations above background noise) in cell cluster 14 (with more than 700 cells); as expected, cells in this cluster expressed TCR Va 7.2 exclusively with no other TCR proteins according to dsb normalization (Fig. 2.4c, Supplementary Fig.

9f). In contrast, the CLR normalized values of the cells in this cluster had higher median values for TCR a/b than TCR-va7.2; both TCRs were similarly distributed and it was thus unclear which was truly expressed given the uncertain noise floor of CLR normalized counts (Fig. 2.4d, Supplementary Fig. 9e). This was also the case for the gamma-delta T cell receptor protein, which was around zero after dsb normalization (Supplementary Fig, 9e-f). CD56, CD3, CD8, and KLRG1 in Cluster 14 were also positive based on dsb (more than 6 standard deviations above background noise) (Fig. 2.4e), thus broadly consistent with the known phenotype of CD8+ MAIT cells[253]. These cells have distinct biological functions from conventional T cells, partly due to their semi-invariant T Cell Receptor (TCR-Va7.2) specific for bacterial metabolic products presented via Major Histocompatibility Complex related protein MR1[254]. Based on CLR normalized protein levels alone, cells in cluster 14 had a phenotype resembling conventional T cells with elevated cytotoxic capacity (TCR a/b, KLRG1 and CD56 positive)[255,256]. Since dsb corrects for protein-specific noise, we hypothesized that the apparent expression of both TCRs in cluster 14 after CLR normalization was likely due to ambient noise present in CLR transformed data. Supporting this notion, the median CLR values (but not the dsb-derived values) were correlated with the 98th percentile values from empty droplets (Pearson correlation 0.8, two-sided p = 2.4e-7, compared Pearson correlation 0.13, two-sided p = 0.39 for dsb), and both the alpha-beta and gamma-delta TCR proteins were among the highest ranked proteins based on expression in empty droplets (Figs. 2.4e-f). To further assess the identity of this cluster, we performed unbiased differential mRNA expression analysis of cluster 14 cells versus other clusters (Fig. 2.4g). Among the top discriminative markers for cluster 14 was the transcription factor ZBTB16 (Fig, 2.4g), which is known to be elevated during iNKT and MAIT cell differentiation[257], expressed by mature MAIT cells[258,259], but suppressed during conventional naïve T cell differentiation[260]. We next constructed a 165-transcript MAIT cell signature derived from the top differentially expressed genes reported in an independent study, which used bulk RNA-seq to compare FACS-sorted TCR-Va7.2+ human MAIT cells versus other T cells lacking this TCR[261]. This MAIT cell signature was significantly enriched (Fig. 2.4h) in differentially expressed genes from cluster 14 (normalized GSEA enrichment score 2.64, p value 1e-10). As this example demonstrates, dsb helped to avoid potential misannotation of a T cell subset and revealed biologically coherent mRNA and protein profiles of MAIT cells. Thus, dsb is

compatible with and can improve downstream analysis outcomes of multimodal single cell data such as TEA-seq.

Figure 4

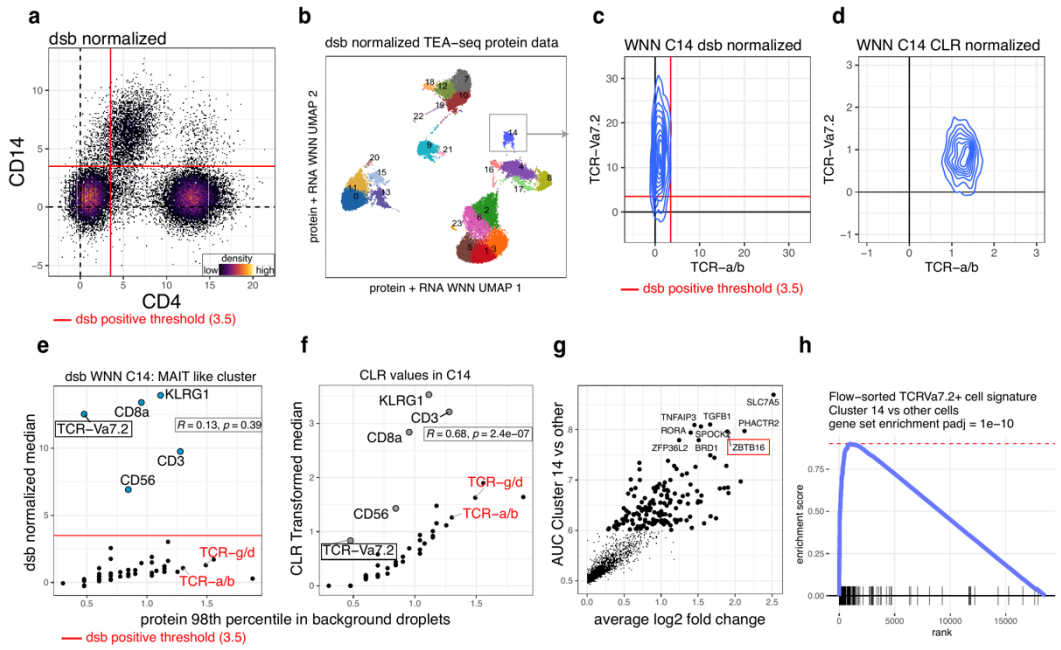


Figure 2.4 Case study II: application of dsb to tri-modal TEA-seq data unmasks a MAIT cell population obscured by noise in CLR normalization.

Analysis of TEA-seq (transcriptome, epitopes and accessibility) tri-modal single cell assay data. **a.** dsb normalization of protein data from TEA-seq showing the distribution of CD4 and CD14 with the same 3.5 threshold used throughout the study. **b.** UMAP plot of single cells and clusters derived by WNN joint mRNA-protein clustering with protein data normalized using dsb. **c.** Bi-axial distribution of the alpha beta and va7.2 T Cell Receptor (TCR) proteins in cluster 14 cells normalized by dsb and **d.** the same cells CLR normalized values. **e.** Similar to Fig. 2.3g but here for cluster 14 from (b) using dsb or **f.** CLR normalized values (y-axis); in both plots Pearson correlation coefficients and p values (two sided) are shown between normalized values (y axis) and values in empty droplets (x axis). **g.** Differential expression analysis (ROC test) of genes in cluster 14 vs other clusters. **h.** Gene set enrichment of a MAIT cell signature constructed from FACS-sorted TCR-va7.2+/MAIT cells compared to other T cells (RNA-seq data from Park *et. al.* 2019*)* with genes ranked by log2 fold change in cluster 14 cells vs other cells as in (g).

2.9 Discussion of Chapter 2 results

Our experiments and computational analyses revealed ambient capture of antibodies by droplets is a major source of protein-specific noise in droplet-based ADT data. Our method, dsb, estimates and corrects this noise component without experimental modifications since we found that it can be reliably estimated using empty droplets,

which are abundant in droplet based single cell datasets. On top of protein-specific noise, cell intrinsic noise was apparent given our observation of the strong correlation (i.e., shared variance) among distinct isotype controls and the average ADT level of background proteins inferred by mixture modeling within single cells. This correlated component affords dsb to implement a conservative approach to estimate and correct for cell-to-cell technical noise, an improvement over approaches that use individual isotype controls or total protein library size because individual variables alone are inherently noisier and could contain more biological (as opposed to technical) signals. We found that application of dsb to both our own and independent multimodal single cell datasets with ADT data improved the identification and annotation of cell types and states based on protein-based or multimodal clustering approaches.

Recent methods proposed to use joint probabilistic modeling of mRNA and protein[242,262] with one of the goals being identification of protein expression above noise. For example, TotalVI[262] uses an mRNA and protein generative neural network model to estimate posterior probability distributions of protein expression, which identified cells with zero, low, or high probability of CD4 protein levels in human PBMCs. As expected, this identified monocytes and T-helper cells based on known low and high surface CD4 protein levels on these cells, respectively; these populations were similarly recovered by dsb normalized populations. While such end-to-end probabilistic models hold promise for single cell analysis, the TotalVI counts are denoised in non-normalized UMI count space–to use these raw UMI counts for downstream analysis tasks outside the probabilistic neural network framework, the values would still need to be normalized, for example via a log transformation. Such probabilistic models are thus complementary to and distinct from dsb, which focuses on low-level protein- and cell-intrinsic denoising and normalization unique to ADT protein data by directly inferring and removing the two noise components detailed in our analyses above. In addition, the specific noise sources revealed by our analyses and approaches to estimate them could lead to more informative prior distributions used by Bayesian probabilistic modeling approaches such as TotalVI. As demonstrated here, the denoised and normalized data from dsb can be used in any downstream analysis application to potentially enhance the results of higher level single cell data analysis methods, such as joint protein-mRNA clustering[251,263–265]

We further detail the experimental evidence for noise sources as well as the modeling assumptions, caveats, and limitations of our method in the Supplementary Note. Briefly, we assessed 1) the robustness of our estimation of protein-specific noise, 2) the sensitivity of dsb normalized values to different methods of defining empty droplets, 3) the impact of different cutoffs for defining background droplets for use with dsb, 4) normalization across batches: normalizing multiple experimental batches together vs. applying normalization separately to each batch, and 5) caveats for using dsb on datasets without isotype control antibody measurements. The use of different methods for defining background droplets had negligible impact on normalized expression values, however, defining a reasonable subset of barcodes as background droplets still requires care. The dsb package documentation provides code to extract and quality control the background droplet population from the raw data matrix. It uses all cell barcodes from the Cell Ranger alignment tool by default, although other alignment tools such as kallisto[266] and CITE-seq-Count[267] can also be used. In our own dataset used above, we also found little differences in dsb-normalized expression values between first merging data across from batches before applying dsb vs. applying dsb to each batch individually. However, in general this could be dependent on the extent of uniformity among the batches. Finally, additional analysis further supported the benefit of including isotype controls to help correct for cell-to-cell technical noise in step II of dsb (see Supplementary Note for details).

The dsb package is computationally efficient and can process on the order of 10^5 cells on a laptop, e.g., the primary dataset in this study (with >53,000 cells) was normalized and denoised in under 4 minutes. The output can be easily integrated with diverse single cell software platforms such as Bioconductor[102], Seurat[268], and Scanpy[269].

2.10 Methods - Chapter 2

The denoised scaled by background normalization (dsb) method

The dsb method is implemented via the R package "dsb" https://cran.r-project.org/package=dsb through a single function call to *DSBNormalizeProtein()*, which models and accounts for 1) protein-specific ambient noise correlated across single cells as captured by average readouts from empty droplets and 2) droplet/cell specific technical noise revealed via the shared variance component associated with isotype control antibodies and background protein counts in each cell. Internally the function is carried out in two major steps. In step I, protein counts in empty droplets are used to estimate the expected ambient background noise for each antibody. Each protein's counts in cell-containing droplets are thus rescaled using this expected noise measurement as:

1. $$\mathrm{Y} = \frac{\log\left(x_i + P\right) - \mu_n}{\sigma_n}$$

Where $\log x_i$ is the natural log of the count for protein Y in cell *i, P* is a pseudocount added to prevent taking the log of zero and to stabilize the variance of small counts, and μ_n and σ_n are the mean and standard deviation of empty droplets for protein Y, respectively, computed in the same way in natural log space with pseudocount *P* added. The value of *P* can be empirically chosen; we use 10 by default, finding this provides good clustering performance and visualization of the CITE-seq data we have analyzed. The transformed expression estimate (*Y*) for the protein in each cell can be interpreted as the number of standard deviations above the expected ambient background noise of that protein. This expression matrix can be returned without further removing technical cell to cell variations in step II, for example if isotype controls are not available, by setting *denoise.counts = FALSE* in the R function, however we strongly recommend using isotype controls and further correcting cell to cell technical variations by fitting and removing each cell's dsb technical component in step II below by setting *denoise.counts = TRUE* and *use.isotype.control = TRUE* (the function default).

In step II, dsb denoises cell-to-cell technical variations by defining and removing the "technical component" of each cell's protein values after ambient correction from step 1. This step fits a model to each cell to learn the background population mean, and then combines this value with the shared variation in values of isotype control proteins. In

the first part of this two-part step, dsb fits a Gaussian mixture model through the expectation-maximization algorithm implemented with the mclust[270] R package to the transformed count of each cell from step 1 with k = 2 mixture components:

2. $$(f x_i) = \phi_1 N_1(x \mid \mu_1, \sigma_1) + \phi_2 N_2(x \mid \mu_2, \sigma_2)$$

In the model above, the log normally distributed proteins of each cell *i* comprising the non-staining noise/background protein subpopulation for that cell are estimated by (N_1), and μ1 is the mean of the background protein subpopulation N_1 in that cell. A noise variable matrix is then constructed by combining all the fitted μ1 values with the isotype control values for all cells. dsb then calculates principal component 1 (i.e. the primary latent component "λ") of these variables in the noise matrix across cells:

3. $$\lambda_1 = \phi_{1,1}(\mu 1) + \phi_{1,2}(Isotype 1) \ldots \phi_{1,p}(Isotype\ p)$$

Where loading vectors in equation III calculated by the R function *prcomp()* are multiplied by the noise matrix, forming each cell's PC1 score λ_1 which determines the cell's "dsb technical component". Finally, the dsb technical component for each cell is then regressed out of the ambient-noise-corrected values "Y" from part 1; the values returned by dsb are the residuals (plus intercept) of a linear model regressing the ambient corrected values on the technical component for each protein. Internally, to implement this step dsb uses a function from the limma[70] package *removeBatchEffect()* for robust and efficient matrix decomposition to fit and then regress out the effect of a specified covariate (in this case the technical component λ_1 from equation 3) from a matrix of variables (proteins) across observations (cells).

We strongly recommend using isotype controls if using the cell to cell denoising step (i.e. if setting *denoise.counts = TRUE*) as we observe that the use of more isotype controls increases the robustness of the calculation of the technical component. See the dsb software documentation on CRAN and the Supplementary Note for additional

information on usage and definition of the technical component in experiments without isotype controls.

CITE-seq on 20 human PBMC samples

CITE-seq data analyzed here were previously used to assess the cellular origin and circuitry of baseline immune signatures[115]; an earlier version of dsb was used therein to normalize the protein data which is identical to the default method implemented in the dsb package, with exception of the pseudocount used (1 vs 10, see below). Experiment details can be found in our prior report[115]. Briefly, oligo-labelled antibodies for sample barcoding (cell "hashing") and surface target protein detection were obtained from Biolegend. After incubating each sample with a barcoding antibody[98], cells from each donor were pooled into one tube and stained with an optimized mixture of oligo-labelled CITE-seq antibodies against target surface proteins. Two experimental batches were performed on consecutive days, using aliquots of the same pool of antibodies for each batch. The pooled donor cells from each of two batches were each distributed evenly across 6 lanes (per batch) of the 10x Genomics Chromium Controller using Single Cell 3' expression reagents (version 2). Sample barcoding (HTO) and target surface protein (ADT) libraries were prepared as in the original CITE-seq report and according to the publicly available CITE-seq protocol (version 2018-02-12, cite-seq.com). cDNA libraries were prepared using the 10x Genomics v2 kit according to manufacturer's instructions. Libraries were sequenced using the Illumina HiSeq 2500 using v4 reagents. We used CITE-seq Count[267] for HTO and ADT read mapping and Cell Ranger for RNA mapping, and cells were then demultiplexed as previously reported[98,99,115] (see Supplementary Note for additional details on demultiplexing, see supplementary Data 1 for a list of antibodies used in this study).

Healthy donor CITE-seq data analysis

Raw CITE-seq data from our prior report[115] were normalized with the dsb package using the default parameters and empty/background droplets as defined by either clear breaks in the protein library size distribution or droplets defined as negative/background by sample demultiplexing with little impact on normalized values (See Supplementary Note and Supplementary Fig. 8). The *denoise.counts* argument was set to *TRUE* which

carries out the recommended step 2 (denoising cell-cell technical variations by estimating and regressing out the technical component for each cell) and the *use.isotype.control* argument set to *TRUE* (defining each cell's technical component by combining isotype control antibody values and the mean of background counts as detailed above). See section below "Assessment of performance of dsb vs CLR" for methods related to normalization comparisons. Uniform Manifold Approximation Projection[271] (UMAP) was run with the umap-learn Python package in R using reticulate with parameters *n.neighbors* = 35, *min.dist* = 0.6. Unsupervised protein based clustering was performed using Seurat[272] to implement the SLM[273] algorithm as we previously reported[115] directly on a distance matrix formed on the protein vs cells matrix of CITE-seq proteins (without isotype controls) after normalizing with dsb (in our original report, using pseudocount 1). We retained these cell type annotations used in the original report but renormalized data for all analysis in this paper using dsb with the current package default pseudocount = 10 which resulted in identically distributed relative protein values across cell clusters (Supplementary Fig. 6c, see also Fig. 5c in Kotliarov *et. al.* 2020).

Assessment of performance of dsb vs CLR

Our CITE-seq PBMC data of ~53,000 cells from healthy donors profiled with 83 phenotyping proteins and 4 isotype controls (as shown in Fig. 1 and 3, from Kotliarov *et. al.*) was used for comparison of CLR and dsb normalization using statistical tests, cell type annotation from protein based clustering and comparison of multimodal mRNA + protein-based clustering. For comparisons, the default implementation of dsb, with *denoise.counts = TRUE* and *use.isotype.control = TRUE*, was compared to the CLR transformation across cells, parameters *normalization.method = CLR* and *margin = 2* in the *NormalizeData()* function in Seurat version 4[251]. The Gap statistic[249] for dsb and CLR normalized data was calculated based on k medoids clustering algorithm with k values from 1 to 20, using with 20 bootstrap samples to obtain the reference null distributions. Differential expression testing of protein markers comparing each cluster to all other clusters was performed for the major cell types in the coarse clustering (clusters C0-C10) as reported in Kotliarov *et. al.*[115] vs all other cells using the *FindMarkers()* function in Seurat to implement a Wilcox test with a log fold change

threshold of 0.3. See section below "Weighted Nearest Neighbor analysis of CITE-seq and TEA-seq data" for information on clustering comparison.

Assessment of dsb on external CITE-seq (protein + mRNA) datasets

Raw and filtered UMI matrices for RNA and ADT counts from Cell Ranger were downloaded from the 10X Genomics website. Background droplets and cells were defined and the default dsb normalization was carried on each dataset as described in the tutorial in the dsb package documentation https://github.com/niaid/dsb. Cells were defined as barcodes in the *filtered* Cell Ranger output, and background drops were defined as after removing the cells from the from the *raw* Cell Ranger output, where a range of ~5e4 to 7e4 background droplets containing protein reads were used to measure ambient background. Background drops could be clearly differentiated from cell containing droplets by an order of magnitude difference in the protein library size distribution (see blue vs orange distributions in Supplementary Figs. 4 a,h,o). The droplets in each of these populations were then subjected to standard scRNAseq quality control metrics based on mRNA content, mitochondrial read proportion and protein library size with filters tuned to each dataset in order to retain only high-quality cells in the cell protein matrix and to remove potential cells from the background protein matrix. The number of cell-containing droplets after QC was consistent with the expected per-lane cell recovery based on the cell loading density of the experiment. Proteins with very low raw data signal (a maximum UMI count < 5 across all cells) were removed prior to normalization, resulting in removal of the CD34 protein from two datasets. After these basic quality control steps, dsb normalization was carried out using default parameters in the dsb package (*denoise.counts = TRUE and use.isotype.control = TRUE*). Cells were clustered on a cell by protein Euclidean distance matrix of dsb normalized values not including isotype control proteins as described above. UMAP was run with *n_neighbors* parameter = 40 and *min_dist* parameter = 0.4. Cluster labels reflect graph-based clustering in Seurat with resolution tuned to each dataset.

Assessment of dsb on external proteogenomic (protein + DNA mutation assay) data

The Mission Bio example data was downloaded from the company's website. Since this dataset only analyzed ten surface proteins, we performed ambient noise removal, rescaling based on counts in the observed empty droplets only (i.e. performing step 1 only by setting the denoise.counts argument to *FALSE*). UMAP was run with the *min_dist* parameter set to 0.4 and the *n_neighbors* argument set to 40 directly on dsb normalized protein values. Clustering was done on a Euclidean distance matrix using Seurat with a resolution parameter set to 0.5 as described above.

Analysis of ASAP-seq (protein + chromatin accessibility) and TEA-seq (protein + mRNA + chromatin accessibility) data

ASAP-seq and TEA-seq data were downloaded from GEO and preprocessed according to the workflows provided in the publicly available analysis code from the original manuscripts. Cell containing droplets were defined as the droplets that passed the authors original quality control metrics. For dsb normalization, we subset non-cells from the raw protein data, estimating noise from the major peak in library size distribution, with quality control to eliminate potential cells from the background matrix, following a similar procedure outlined in the dsb documentation with some modification for ASAP-seq data where mRNA data are not available thus background was estimated based on protein alone from the subset of droplets that did not pass the authors quality control metrics for cells. For comparison, in both datasets cells were normalized with CLR (across cells, margin = 2 in the *NormalizeData()* function using Seurat) and for TEA-seq, an additional log transformation with library size scaling factors (*NormalizeData()* function with parameter *normalization.method = "LogNormalize"*). Analysis of differentially expressed genes in specific clusters vs all other cells was carried out in Seurat with the function *FindMarkers()* using an ROC test. Gene Set enrichment analysis of the MAIT cell signature was performed with the fgsea package[274] based on genes ranked by the log2 fold change of genes in cluster 14.

Weighted Nearest Neighbor analysis of CITE-seq and TEA-seq data

For multimodal clustering of the TEA-seq and CITE-seq healthy donor datasets, we used the weighted nearest neighbor algorithm[251] with the Seurat function *FindMultimodalNeighbors()*, with slight modification. In pilot analysis of the WNN

algorithm we found both CLR and dsb joint embeddings and clustering improved by using protein data directly instead of compressing protein data into principal components. We used the normalized values of 45 (TEA-seq) and 69 (our healthy donor CITE-seq data) phenotyping proteins directly in the joint model (we first removed 14 uninformative / poor performing proteins that had very low average dsb values across all protein based clusters from the protein data matrix from the healthy donor CITE-seq data). We compared the same joint clustering approach with the only difference being the normalization used in the input data. In analysis of both the CITE-seq and TEA-seq datasets, the mRNA data was compressed into 30 principal components; the same 30 mRNA principal components were combined with either dsb or CLR normalized protein data for joint clustering. First, mRNA data were normalized with the Seurat function *NormalizeData()* with the parameter *normalization.method = "LogNormalize",* implementing a natural log transformation, standardizing by the library size and multiplying values by 1e4. These values were compressed into 30 principal components based on scaled values for variable genes selected by setting the *FindVariableFeatures()* function with the *selection.method* parameter set to 'vst'. For the CLR WNN model, protein data were normalized by the CLR across cells (using the Seurat function *NormalizeData()* with normalization.method = "CLR" and margin = 2). For the dsb WNN model, data were normalized using the default implementation of dsb, (parameters *denoise.counts = TRUE, use.isotype.control = TRUE*). Seurat was then used to separately cluster the two weighted nearest neighbor graphs constructed from mRNA principal components and either CLR or dsb normalized input protein data.

Data Availability

Datasets used in this analysis are available to download at: https://doi.org/10.35092/yhjc.13370915. The public datasets included in the data repository were downloaded online and are also available from 10X genomics at https://support.10xgenomics.com/single-cell-gene-expression/datasets and from Mission Bio at https://missionbio.com/capabilities/dna-protein/#Data. ASAP-seq and TEA-seq datasets were downloaded from GSE156477 and GSE158013 respectively.

Code Availability

The dsb software package is available for download on CRAN:

https://cran.r-project.org/package=dsb.

An analysis workflow with R code to reproduce the analysis results reported in this manuscript are available for download from github:
https://github.com/niaid/dsb_manuscript/.

3 MULTISCALE DECONSTRUCTION OF VACCINATION RESPONSES REVEALS HIGH ANTIBODY RESPONDERS TO UNADJUVANTED VACCINES ARE NATURALLY ADJUVANTED

3.1 Abstract

We developed a multiscale analytical framework to deconvolve human population, vaccine formulation, immune cell subset, and single cell variations in response to vaccination using multimodal single cell data. Integrating mixed effects models with computational reconstructions of cell state deconvolved past bulk derived “signatures” of response and further defined classes of shared and cell type specific perturbation states across biological scales. A contrast approach identified AS03 adjuvant specific perturbations including B lymphocyte survival and sensory receptors unrestricted to a particular pathogen class. Remarkably, monocytes and dendritic cells of high responders appeared “naturally adjuvanted” with baseline elevation of the same sensory phenotypes which were induced post vaccination specifically by adjuvant AS03. This adjuvant like high responder baseline setpoint reflected tightly coupled multicellular transcriptional circuitries including correlated sensory receptors and interferon signaling. Primed baseline circuitries themselves were then coherently induced 1 day post vaccination within the same cell types following influenza vaccination and by mRNA vaccination in a separate cohort. The naturally adjuvanted transcriptional state also extended to cell intrinsic differences in phosphoprotein signaling competence

following *ex vivo* stimulation with PRR ligands. Together, discovery of a naturally adjuvanted immune setpoint and mapping of detailed circuitries open avenues for immune response engineering, while our analytical approach provides a framework for future systems biology studies utilizing population variation and multiscale modeling to understand human in vivo perturbation phenotypes.

3.2 Introduction

Human immune systems exhibit substantial person to person variation[34,40,275]. Population variations in immune responses outcomes to the same perturbation, such as antibody responses to vaccination, can be linked to cellular and molecular immune system components using top down systems biology approaches[20]. Such studies have used unbiased profiling to identify signatures of timed perturbation states and quantitative antibody response[39,143,159,160,162,163,187,194,276], including those mediated through individual intrinsic factors, such as genetics[134] age[161,164], and sex[145]. Furthermore, accumulating evidence from these studies supports the hypothesis that immune system status prior to perturbation can influence response quality[39,161,223,277,278]. For example, we identified bulk transcriptome signatures reflecting a stable immune system "setpoint" linked both to improved antibody response following vaccination in healthy individuals and to plasmablast activity during disease flares in lupus patients[115]. More recently, bulk blood transcriptome profiling studies identified prognostic signatures in healthy children at risk of type 1 diabetes prior to development of the disease[279], and at baseline in cancer patients prior to immunotherapy induced autoimmunity[92].

These systems immunology approaches endeavor to develop a holistic understanding of immune cell processes which elicit optimal immune responses[33,280]. Several technical challenges impede moving from biomarker signatures identified to date into such an integrated picture. Protein phenotypes measured using cytometry cannot assess diverse internal cell states captured by transcriptomics, however bulk blood transcriptome profiles are confounded by substantial between-individual variations in circulating immune cell subset frequency[9,39,117]. Single cell transcriptomics can further resolve cell states, yet interpretation remains challenging when measuring chromatin accessibility or

mRNA alone given existing knowledge cataloging immune cell types using surface protein[38,39,117,281]. Multi-modal single cell transcriptome and protein profiling methods such as CITE-seq[44] hold promise for unifying these modalities to derive more interpretable insight from immune system profiling. However, how to model timed perturbation responses and define meaningful variations spanning biological scales from individuals, to cell types and single cells remains a major unmet analytical challenge.

We profiled 52 PBMC samples from 26 individuals before and after vaccination with pandemic influenza vaccines using CITE-seq. Individuals were nested into three groups: those with 1) high or 2) low antibody responses to an unadjuvanted vaccine and 3) individuals vaccinated with an AS03 adjuvanted vaccine. In this work we developed a multilevel modeling framework to integrate population variations, response kinetics and reconstruction of single cell states. These approaches deconvolved cell phenotypes linked to timed vaccine perturbations and further revealed phenotypes associated with desirable emergent response properties induced specifically by AS03[282]. In addition, we found extensive rewiring of baseline cellular innate cell circuitries of high compared to low antibody responders to unadjuvanted vaccination which reflected elevated innate immune cell potential. Comparative analysis of cell states induced specifically by adjuvant AS03 with these unadjuvanted baseline cell phenotypes, and further experiments measuring early phosphoprotein signaling responses to pattern recognition receptor stimulation, revealed high responders to the unadjuvanted vaccine had a "naturally adjuvanted" baseline immune set point. Our analytical approach paves the way for multiscale analysis of timed perturbation studies using single cell data in humans. Furthermore, our findings open avenues for defining targets of immune response engineering and vaccine development based on defining population variation in precise cellular phenotypes linked to desirable immune responses.

3.3 Results

3.4 CITE-seq experiment design to measure human response variations to timed vaccine perturbation across biological scales

We assessed 52 PBMC samples from n=26 donors pre and post vaccination using CITE-seq. Subjects received either the 2009 seasonal + pandemic type A strain vaccine,

or a pandemic avian influenza strain formulated with oil in emulsion adjuvant AS03. Twenty subjects including n=10 high and n=10 low responders in the unadjuvanted vaccine group were profiled at baseline and on day 1 or 7 post vaccination (Fig. 3.1 a,b). We derived sources of noise in CITE-seq protein data and developed a dedicated normalization method[100], then extensively tested the reliability of CITE-seq to recover and unify known cell surface and transcriptome phenotypes. For example, we gated activated B cells vs plasmablasts (which are prone to experimental loss) based on CITE-seq surface protein levels which recovered cell type specific transcriptional signatures[283] derived from the same flow sorted subsets (Fig. 3.2a).

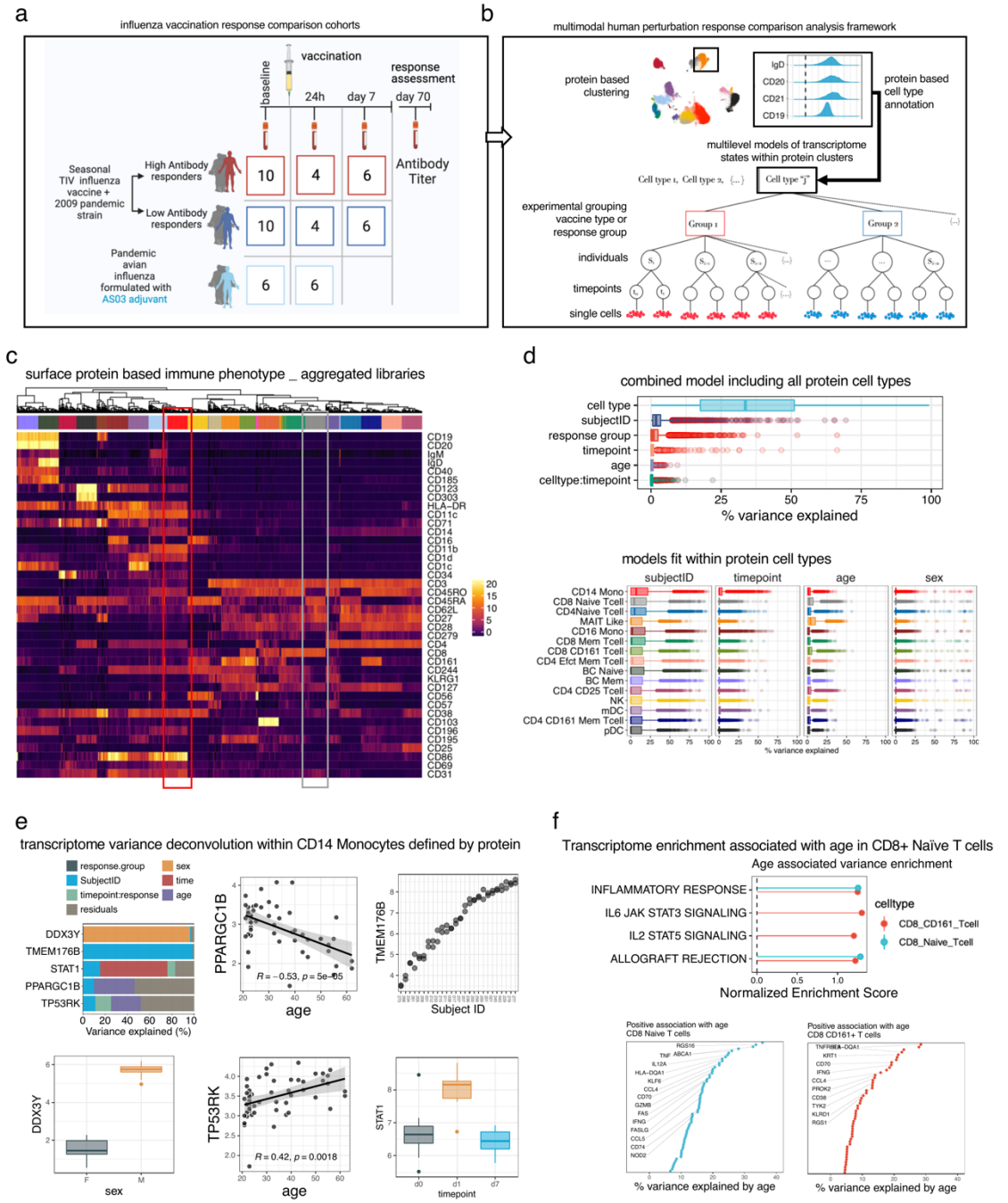


Figure 3.1 Single cell portraits of human vaccination response through within cluster mixed models comparing vaccination effects over time between groups

a. Human vaccination response study outline; CITE-seq data was generated from n=52 PBMC matched pre- and post-vaccination PBMC samples from n=26 subjects including 2 response groups and two vaccine formulations. 10 high and 10 low responders from the 2009 TIV + pandemic H1N1 influenza vaccination without adjuvant were profiled with a subset of 8 and 12 subjects split evenly between high and low responders profiled on day 1 and 7 respectively. 6 subjects vaccinated with a pandemic H5N1 avian influenza vaccine formulated with adjuvant AS03 were profiled at baseline and day 1 post vaccination. **b.** The hierarchical structure of the CITE-seq data collected on 52 PBMC samples for a single cluster is shown to motivate necessity of multilevel modeling approach for transcriptome analysis. Clusters are based on surface protein (select proteins from naïve B cell cluster shown); within each cluster modeled with weighted mixed effects models clusters are represented by cells from PBMC samples indexed by individual, timepoint and different response groups (high and low responders) and vaccine group (unadjuvanted vs adjuvanted). **c.** For each of > 700 samples aggregated by protein based cell type and individual x timepoint, the average dsb normalized protein expression in each cell type is shown. **d.** Top: the fraction of variance explained in a multivariate model across libraries aggregated by cell type, individual and timepoint; bottom: as in d, with models fit within each protein based cell type, i.e. within colored columns of c. **e.** variance fractions for 5 genes within CD14 monocytes (model from bottom of d) with additional visualizations of gene expression (y axis) vs the experimental factor (x axis) explaining maximal variance for the 5 genes. **f.** top: enrichment of genes ranked by their variance explained by age; subset of genes with positive association with age in CD8 naïve and CD161+ T cell clusters; bottom: select genes positively associated with age within the two cell types.

3.5 Transcriptome variation decomposition into protein based cell type, individual, age, sex and vaccination effects

Cells clustered based on dsb normalized and denoised[100] level of 82 CITE-seq surface proteins enriched known immune phenotypes (Fig. 3.2 c, d). Individuals were represented in a majority of clusters at both timepoints (Fig. 3.2 e,f). Certain subsets represented by only two to three subjects (e.g. NKT and CD57+ CD4 T cells) were detected within individuals at both timepoints as expected based on stable, within-individual longitudinal variation in human cell phenotypes[39,281] (Fig. 3.2 f). We next deconstructed variation of each gene into that attributable to cell types, individuals, intrinsic factors (age, sex) and vaccination using multivariate mixed effects models. Models first fit to each gene across more than 700 transcriptome libraries indexed by cell type, individual, and timepoint (Fig. 3.1c, columns), intuitively revealed cell type (Fig. 3.1d, top) explained on average more than 30% of gene variation across the

transcriptome (range 0-100%), consistent with known cell type specific transcriptome profiles. To further identify intrinsic and vaccination effects independent of cell type specific expression effects, we next fit models within protein based subsets (Fig. 3.1d, bottom). (Fig. 3.1e) highlights variance fractions for select genes within CD14 Monocytes. As expected sex explained nearly all variation in expression of Y-chromosome gene DDX3Y. More unexpectedly, a transcription factor genetically linked to rheumatological pathology[284], PPARGC1B, and apoptosis regulator TP53RK were negatively and positively associated with age respectively. Globally, models identified substantial between-subject variations (Fig. 3.1D) which, for example, accounted for nearly 100% of variation in TMEM176B, an inflammasome signaling regulator[285]. Finally, timepoint relative to vaccination accounted for more than 50% of variation in STAT1; related differential expression models revealed vaccination induced expression of this gene 24h post vaccination (see below). Age contributed high variance fractions across genes in CD8 naïve and CD8+ CD161+ T cells relative to other cell types; inflammatory processes were enriched among genes positively correlated with age (Fig. 3.1f), consistent with sterile inflammation linked to aging[155].

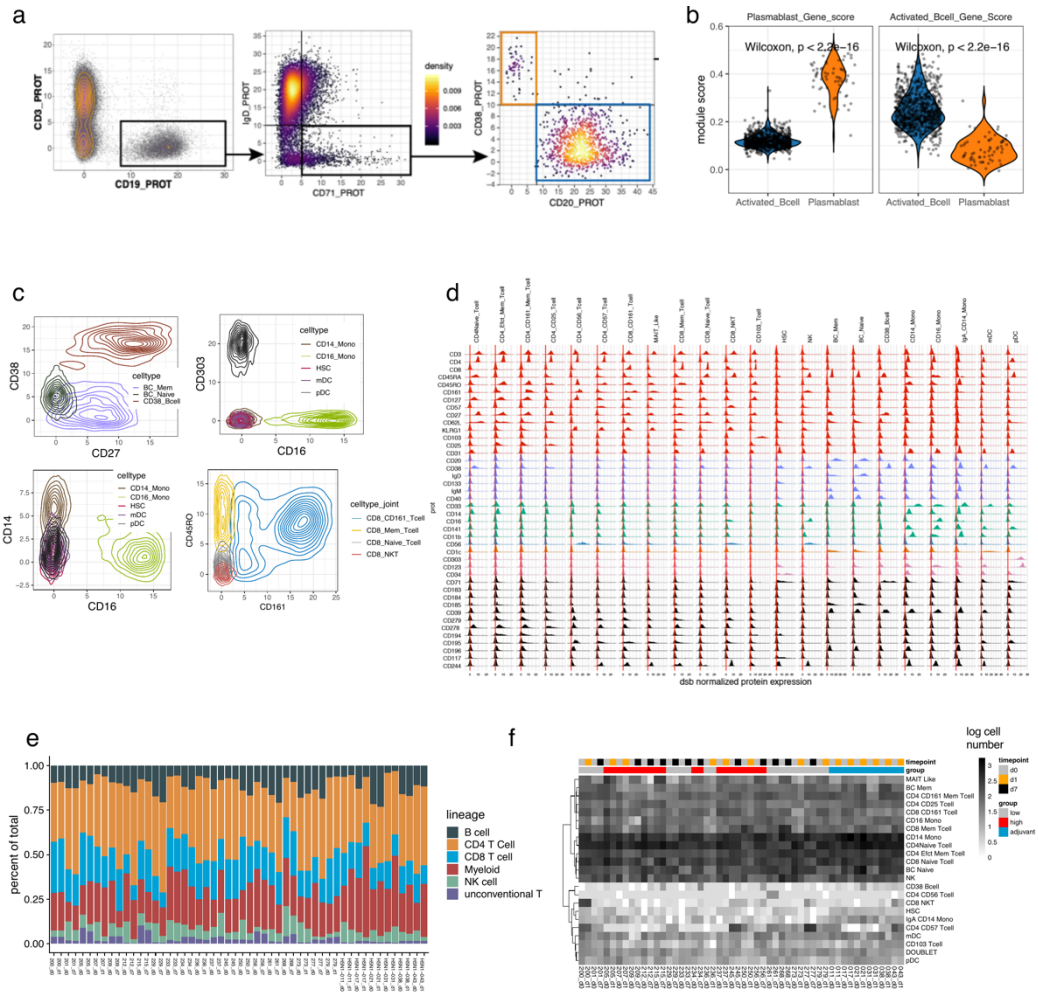


Figure 3.2 Quality control of CITE-seq perturbation transcriptome response detection and surface protein phenotypes compared to microarray and flow cytometry data

a. Manually gated cell populations based on dsb normalized CITE-seq surface protein expression, orange box: plasmablast (CD19+CD71+IgD-CD20-CD38++) and blue box: activated B cell (CD19+CD71+IgD-CD20+CD38+/-). **b.** Transcriptome analysis of gene module scores specific to each gated populations (as in Ellebedy et. al.) p-values = unpaired Wilcoxon test between populations. **c.** Density distribution of dsb normalized protein expression binned by protein based cluster for select populations. **d.** Hierarchically clustered histograms of dsb normalized protein distribution within each protein-based cluster. A select subset of proteins are shown and are colored by the main cell populations that they are most informative for discriminating. Red = T cell proteins, light blue = B cell proteins, green = monocyte proteins, dark blue = NK cell proteins, orange = pDC proteins, pink = pDC/HSC markers, black = cell state markers. **e.** The percentage of total cells for each PBMC sample in each major lineage black = B cell, orange = CD4 T cells, blue = CD8 T cells, red = myeloid (all monocytes, HSC, mDC and pDC), green = NK cells, light grey = unconventional T cells (MAIT-like and CD103 + T cells). **f.** Log number of cells per sample by protein based cluster. Individual specific proteins are detected at both timepoints.

3.6 Bulk day 7 transcriptional correlates of antibody response are derived from a small population of plasmablast cells and not naïve or memory B cells

We next used similar mixed effects models to define genes coherently perturbed across individuals by vaccination on days 1 and 7 after adjusting for variation in donor expression, age, sex, technical factors and baseline antibody titers (see methods). Enrichment based on genes ranked by day 7 vaccination effect size from the non-adjuvanted influenza vaccine revealed naïve B cell CD4 memory T cell metabolic processes and activation (Fig 3.3a,b). The proportional shift circulating plasmablasts are hypothesized to drive known day 7 blood transcriptome signatures predictive of antibody response to multiple vaccines[39,163,276]. Indeed, plasmablasts expressed the highest levels of predictive day 7 bulk transcriptome signatures relative to other subsets (Fig. 3.3c). B cell maturation antigen (BCMA) receptor TNFRSF17, had the highest fold change in both microarray and aggregated CITE-seq data (Fig. 3.3d). Absolute deconvolution of reads to each cell type revealed nearly all of the TNFRSF17 UMI counts were derived from just n=89 day 7 CD38high CD20- plasmablast cells and not naïve or memory B cell subsets (Fig. 3.3e).

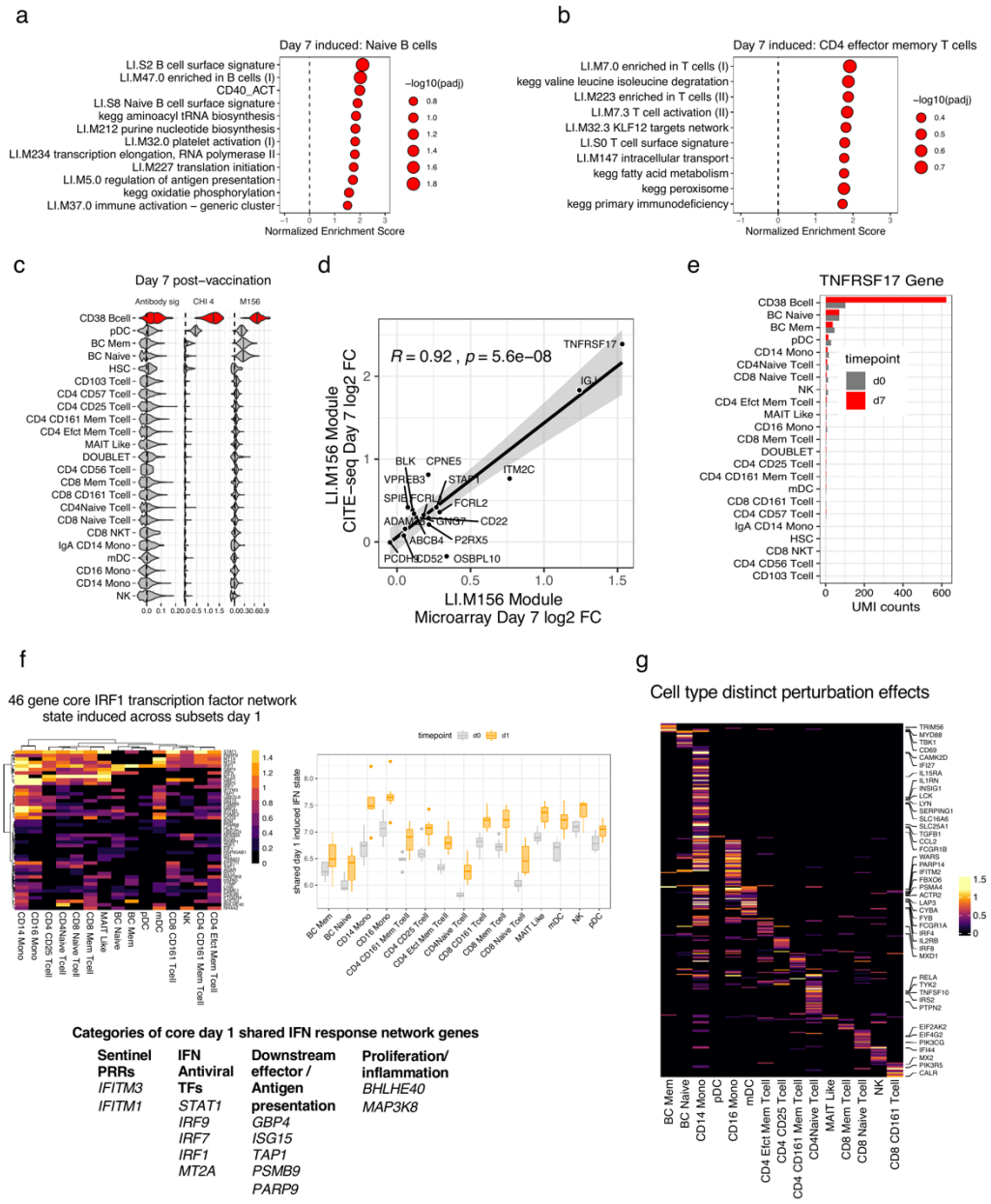


Figure 3.3 Deconvolution day 7 antibody titer associated transcriptome signatures and additional shared and cell type specific early day 1 perturbation phenotypes.

a. Perturbation phenotypes of naïve B cells day 7 post vaccination. Gene set enrichment based on model adjusted post vaccination effect size. **b.** As in a, for memory CD4 T cells. **c.** Protein based cell type specificity of day-7 bulk transcriptomic based gene expression signatures predictive of antibody response from previous systems biology studies of influenza vaccination (Supplementary table 1). Single cell level module score distribution shown for day 7 cells for each cell type. **d.** Correlation between genes in M156 detected in CITE-seq (sample level pseudobulk) vs microarray data (Pearson correlation). **e.** Composition of raw counts of the B cell growth factor receptor TNFRSF17 gene, a driver of M156 on day 7 across protein-based cell types (shown in (e)) shows CD38++ B cells (plasmablasts) are the primary source of the signal. **f.** Left: Heatmap of model adjusted log fold change 24h post vaccination vs baseline of a core interferon signature shared across subsets – genes selected were increased in at least 5 subsets with logFC > 0.1 and raw p value < 0.05. Right: the average expression of the

core shared interferon signature genes across subsets. **g.** As in f, highlighting genes more specifically induced within a single cell type post vaccination.

Deconvolution of the early response to unadjuvanted influenza vaccination reveals shared and cell type specific patterns

Unadjuvanted influenza vaccination response studies consistently report interferon stimulated gene expression (ISG) detected early (1-3 days) post vaccination in bulk transcriptome data. Furthermore, ISG and antigen presentation gene upregulation on day 1 has been found to correlate with higher antibody response[187] though cellular origins of the response genes were unknown. Based on microarray profiling 4 select subsets, early reports hypothesized this signal derived primarily from DCs on day 3[188] or monocyte/granulocytes on day 1[194] however, unbiased profiles of all cells mediating these responses are uncharacterized. Here, unbiased CITE-seq assessment and enrichment against curated pathways, including influenza vaccine signatures curated from the literature, identified three broad patterns of coherent cell perturbation phenotypes 24 hours following vaccination based on their localization across cell subsets (Fig. 3.4a).

The first pattern was characterized by significant enrichment of genes downstream of type I and type II interferon signaling shared across cell types (Fig. 3.4a). 46 “core genes” collectively induced within least 5 cell types each (Fig 3.3 f, g), captured this shared state including transcription factors IRF1 (notably, induced across 15 cell types) STAT1, IRF7, and IRF9, pattern recognition receptor (PRR) genes IFITM1 and IFITM3, inhibitors of vial transcription GBP1[286] and ISG15[287], and antigen presentation genes TAP1, and PSMB9 (Fig. 3.3 f,g). The second pattern included states unique to classical non nonclassical monocytes, such as adhesion molecule ICAM1, JAK2, antigen presentation / HLA genes, and inhibitors of viral replication OAS3[288], and ISG20[289]. The third pattern represented cell type specific perturbation genes (Fig. 3.3 f,g), most notably, inflammatory processes uniquely induced within classical monocytes. The “reactome interferon signaling” pathway (Fig. 3.4a) reflected all three response patterns, with 10-15 shared ISGs across subsets, specific ISGs shared by classical and non-classical monocytes, and a set of classical monocyte specific genes (Fig. 3.4b). Normalized expression of genes driving this pathway within classical monocytes clustered individual samples distinctly by time relative to vaccination,

indicating coordinated cell perturbation phenotype across individuals, as intended by our mixed model framework (Fig. 3.4c).

Genes driving the classical monocyte "IL6 production" pathway reflected early initiators of inflammation MYD88, DDX-58 (RIG-I), TNF and TRAF6. Inflammatory processes were further implicated by monocyte specific expression of IL-1 and IL-15, and chemokine CCL2[290] (Fig. 3.3g). Classical monocytes were also enriched for hypoxia and mTORC1 signaling pathways (Fig. 3.4a). While live influenza can activate and subvert mTOR to support viral replication[291]; inactivated vaccination more likely reflected the role of mTOR in supporting inflammation[292]. Analysis of genes driving the pathway suggested mTOR induced glycolytic metabolism, a process induced after VZV vaccination[204] and linked to non-specific innate memory in monocytes[293]. mTOR enrichment within CD25+ CD4 effector T cells, MAIT-like cells, mDCs and NK cells may have been intrinsically induced by TIV or by monocyte specific expression of IL-15 (Fig. 3.3g), an essential cytokine for activating mTOR in human NK cells[294].

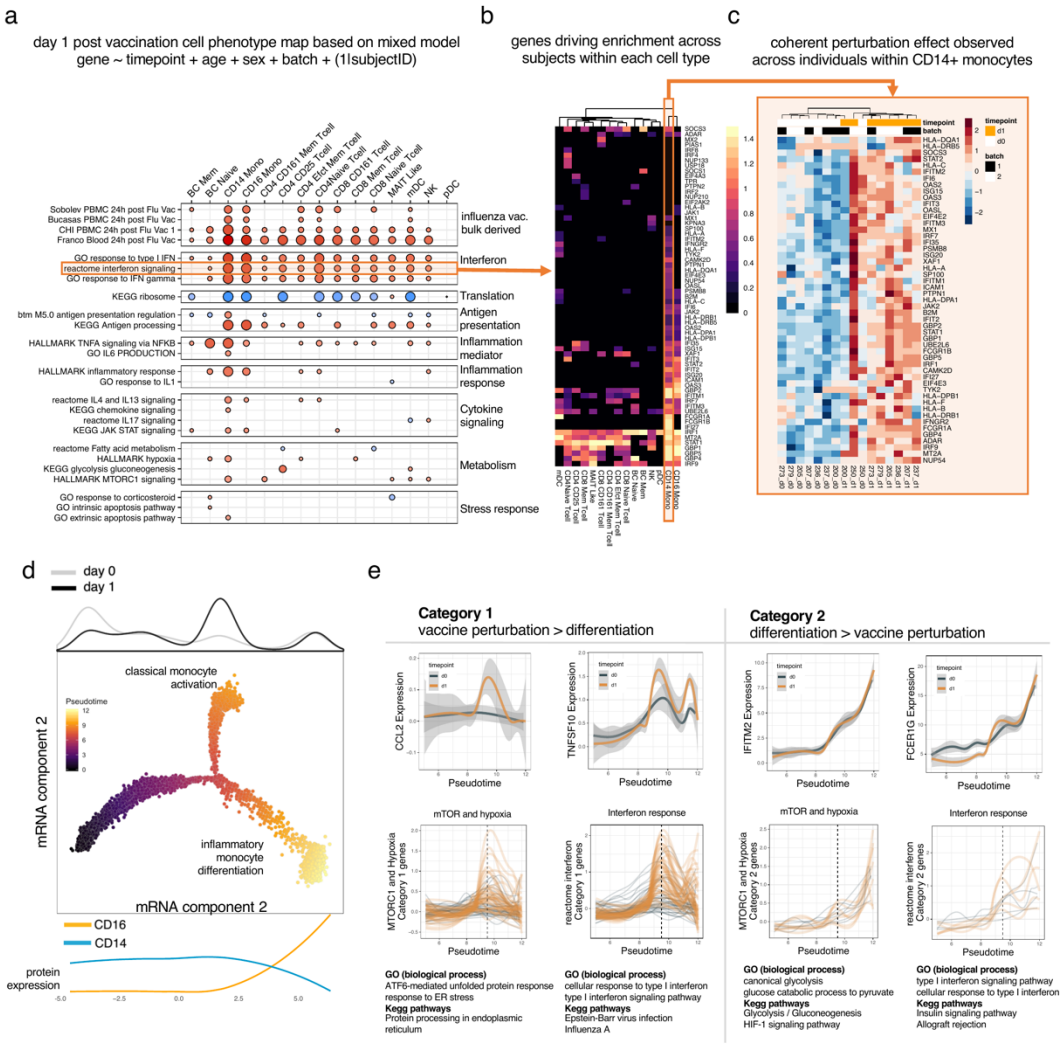


Figure 3.4 Deconstruction of transcriptome perturbation phenotypes induced day 1 post vaccination with seasonal TIV + 2009 pandemic strain vaccine

a. Day 1 post vaccination transcriptional response within protein-based cell types. Gene set enrichment (red = positive enrichment/upregulation, blue = negative enrichment/ downregulation) of modules based on genes ranked by pseudobulk weighted linear mixed effects model baseline vs day 1 effect size (see methods). Interferon response modules tend to be increased across cell types. **b.** Leading edge genes from the reactome interferon signaling module across subsets with enrichment adjusted p values < 0.05. **c.** The leading edge genes within the CD14 monocyte cluster in the reactome "interferon signaling" module demonstrate a coordinated change of IFN genes in monocytes post vaccination. **d.** Pseudotime trajectory inferred with the DDR-tree algorithm constructed with genes changing across monocyte subsets 24h post vaccination (see methods). The true timepoint relative to vaccination of each cell along mRNA trajectory component 1 is highlighted in the top marginal histogram; cells are colored by inferred pseudotime. Three branches from left to right are enriched for resting classical monocytes, activated classical monocytes from post vaccination, and nonclassical monocytes. Cells progressively downregulate CD14 and upregulate CD16 protein level along the rightmost branch; protein data shown in the bottom margin basis spline fit to dsb normalized protein level for CD14 and CD16 (protein levels were not used to construct the trajectory). **e.** Gene expression of select leading edge genes from

enrichments in CD14 monocytes based on branch-dependent differential expression show two broad patterns. Pattern 1 genes are perturbed by vaccination with highest expression in post vaccination classical monocytes – dashed line at pseudotime value of 9.5 represents the peak of activation. Pattern 2 genes continuously increase across pseudotime and have highest expression in CD16+ cD14- non-classical monocytes. The top row shows example genes from each category. The bottom row shows the subset of genes falling into each category from the combined hallmark MTORC1 signaling/Hypoxia pathways and reactome interferon signaling pathways. Below each category / pathway, enrichment of gene ontology (GO) biological process and Kegg pathways for the subset of genes from each pathway and category.

3.7 Multiscale subset and single cell reconstruction with protein integration resolves interwoven monocyte perturbation and differentiation states

We next explored how results from these statistical models could be coupled to unbiased single cell computational reconstructions of transcriptome states. Embedding monocytes in a tree based[295] latent space identified a three branched mRNA-based computational reconstruction of pseudotime[296]. Inferring quantitative relationships between cells in such reconstructions can be error prone[297,298], however, a multiscale approach integrating 1. Cell subset level statistical results detailed above, 2. Time kinetics relative to vaccination, and 3. protein information from the same cells, revealed finer shades of monocyte phenotypic variation in response to vaccination. The ends of the three branches were enriched with pre-vaccination classical monocytes (left branch), day 1 post-vaccination classical monocytes (top branch) nonclassical monocytes present equally pre and post vaccination (right branch) (Fig. 3.4d top). The canonical differentiation process of classical to non-classical monocytes is defined by loss of CD14 and gain of CD16 protein; CITE-seq protein levels were thus critical for identifying the third branch as a differentiation process based on decreasing CD14 and increasing CD16 levels (Fig. 3.4d, bottom margin). Integrating the monocyte genes coherently perturbed by vaccination (leading edge genes from Fig. 3.4a) at the subset level with this single cell computational reconstruction identified two categories of genes based on branch-dependent differential expression (see methods). Category 1 genes, had perturbation effects within either CD14 monocytes alone (CCL2, defined above as a monocyte-specific perturbation gene) or both within CD14 and CD16 monocytes (TNFSF10), whereas category 2 genes (e.g. IFITM2, FCERG1)

continuously increased across the spectrum of pseudotime with highest expression in nonclassical monocytes (Fig. 3.4e, top row). We further investigated the IFN response and mTOR / hypoxia perturbations within classical monocytes at the single cell level by testing genes driving those pathways for branch dependent differential expression. IFN response genes mostly fell in category 1 (more than 40 genes) though 5 genes, PTPN1, IFITM2, IFITM3, HLA-C and EIF4E2 followed category 2. The combined mTOR and hypoxia pathway genes followed a similar pattern, though notably the genes falling in category 2 were more enriched for glycolysis than pattern 1 genes which were enriched for ER stress (Fig. 3.4e, bottom). This integrative multiscale analysis thus suggests the glycolytic shift in classical monocytes in part captures a process where classical monocytes acquire a more inflammatory monocyte-like state, highlighting interwoven activation and differentiation processes.

3.8 Vaccination with adjuvant AS03 induces a pattern recognition sentinel state unrestricted to pathogen class

Through an unknown mechanism, AS03 enhances vaccine potency by eliciting increased level and diversity of anti-influenza antibodies compared to unadjuvanted vaccines, even when formulated with $1/10^{th}$ the antigen dose[282]. Previous studies of the cellular response to AS03 have described strong early ISGs in innate cells[217–219,282] induced by this adjuvant by comparing the dose-sparing AS03 formulation a low dose control formulated with PBS. Here, we used a different approach in order to define AS03-specific stimulation phenotypes beyond early ISG and inflammatory activation. We first applied a statistical contrast defining the difference in the 24h fold change between the AS03 adjuvanted vaccine versus the unadjuvanted vaccine described above (formulated at therapeutic dose for an unadjuvanted vaccine). We next validated the AS03-specificity and strain independence of these phenotypes using data from an external study which profiled FACS sorted immune lineages (e.g. total B cells, T cells) from subjects receiving the same vaccine formulated with AS03 vs PBS[217] (Fig. 3.5a). By design, this approach identified only processes which were perturbed to a statistically greater degree or were specific to AS03 including positive enrichment of several pathways related to surface receptors in monocyte and mDCs (Fig 3.5b, red) which were highly concordant within analogous more coarsely defined innate subsets in the validation cohort (Fig. 3.5b, light blue). Further investigation of the leading edge

genes driving these enrichments revealed they were related to innate sensory capacity unrestricted to a particular pathogen class. In CD14 monocytes, AS03 specific sensory genes included FPR2, which enhances immune cell chemotaxis in response to bacterial metabolites[299] and c-GAS, a cytosolic DNA sensor which activates antiviral response via STING[300] (Fig. 3.5c). Within mDCs, inflammatory chemotaxis receptors FPR1[299] and CCR1[301] were specifically induced in subjects receiving the AS03 formulation. P2RY13 an ADP sensor active during inflammation[302], and TLR4, the PRR for bacterial lipopolysaccharide[303] were uniquely induced by AS03 in both subsets (Figs. 3.5 c-d). To examine variations in these phenotypes at the level of single cells we sub-clustered monocytes and mDCs together based on mRNA and protein[251] (Fig. 3.5e). A mixed effects count model identified one sub-cluster "C2" increased in frequency across individuals on day 1 (Fig. 3.5f). Relative to other clusters, day 1 cells from C2 had an activated "DC-like" transcriptome with high expression of HLA genes and IFN inducible transcripts (Fig. 3.5g) although mDCs clustered independently in C5 likely driven by their distinct protein phenotype. The fraction of cells expressing both sentinel genes and interferon response genes were nearly uniformly increased early after vaccination across sub-clusters (Fig. 3.5h). At the module level, all sub-clusters increased expression of AS03-specific sentinel phenotypes including sub-clusters that decreased in frequency post-vaccination (Fig. 3.6a). The activated C2 monocyte phenotype thus emerged after vaccination, yet all sub clusters had a "responsive" internal state demonstrated by their upregulation of both AS03 specific and interferon response states.

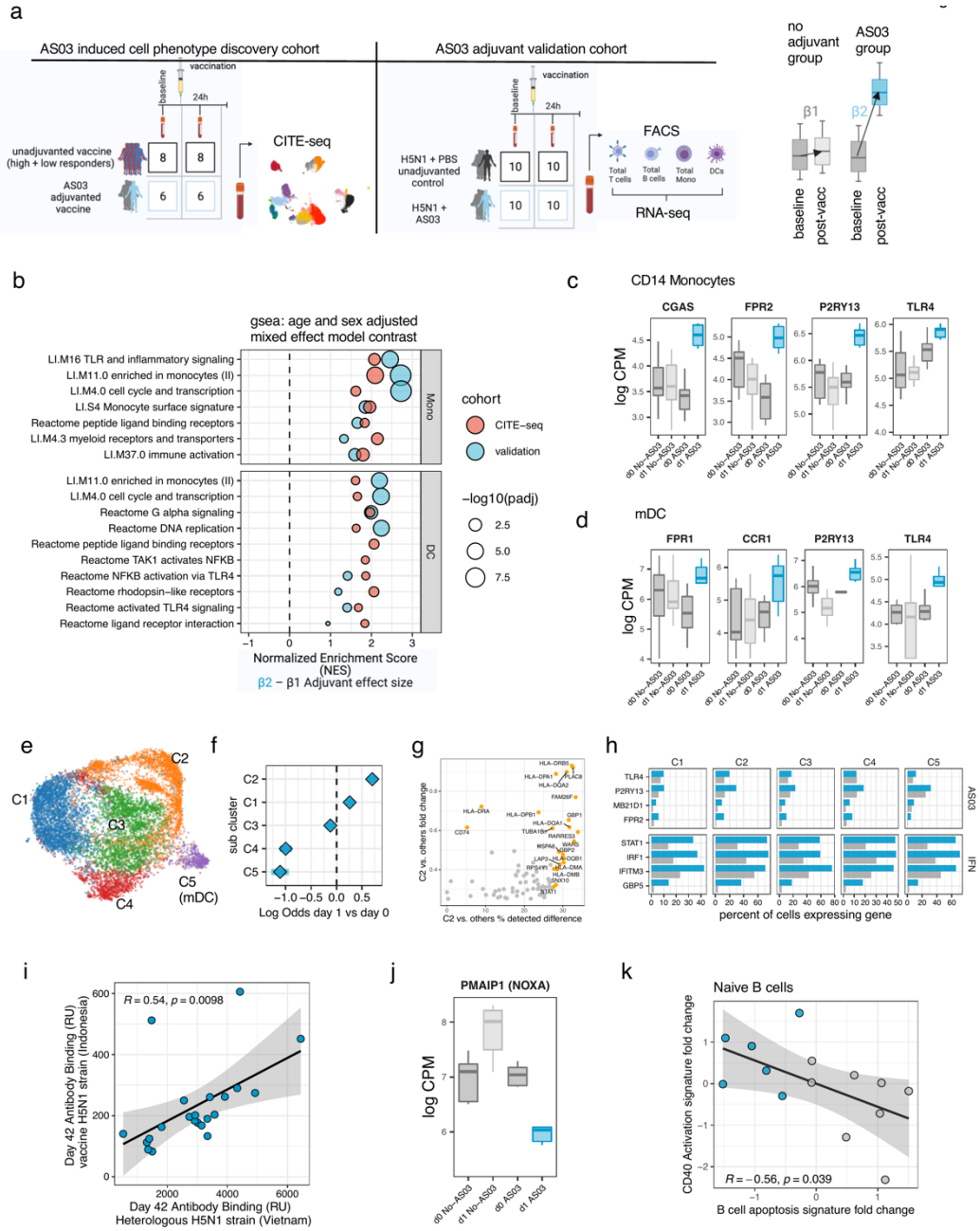


Figure 3.5 Early transcriptional responses to AS03 adjuvanted vs non-adjuvanted vaccines

a. Schematic to illustrate the contrast applied within protein based clusters to dissect AS03 adjuvant specific perturbation phenotypes. Right: the model contrasts within each cell type the difference in fold change between AS03 adjuvanted and unadjuvanted subjects, as shown in the schematic with boxplots. Genes are then ranked for enrichment based on the effect size of this contrast reflecting AS03 specificity, e.g. modules with positive normalized enrichment score have higher day 1 fold change in the AS03 vaccine group compared to the unadjuvanted vaccine. **b.** Gene set enrichment analysis of weighted pseudobulk mixed model comparing transcriptional response 24-hours post vaccination between AS03+H5N1 vs. H1N1 non adjuvated vaccine in Classical Monocytes and mDCs. **c.** Select genes driving difference in perturbation response distinct to AS03 adjuvant within CD14 monocytes and **d.** in mDCs. **e.** Sub-clustering of CD14 monocytes and DCs UMAP plot labeled by cluster. **f.** Mixed effect

binomial count model comparing baseline to day 1 post vaccination cluster cell frequency as a fraction total monocytes, x-axis shows the log odds of cells from a cluster being from day 1 vs day 0, individual modeled with random effect. **g.** Top discriminative genes from cluster 2 relative to other clusters x-axis is the average difference in the fraction of cells expressing the gene from C2 vs all other clusters; y axis–log2 fold change C2 vs all other clusters. Tests done on day 1 cells only. **h.** fraction of cells expressing genes at baseline (grey) and day 1 post vaccination (blue). **i.** The correlation between antibody avidity to the heterologous strain (x-axis – H5N1 Vietnam HA) vs the vaccine strain (y-axis – Indonesia H5N1 HA) (Pearson correlation) measured by surface plasmon resonance assay on day 42 post vaccination in subjects receiving AS03 adjuvant. **j.** PMAIP1 (NOXA) expression across donors within naïve B cells pre and post vaccination. **k.** correlation between the day 1 fold change in the CD40 activation score and the apoptosis signature in naive B cells.

3.9 AS03 adjuvant induces an apoptosis suppression and survival state in lymphocytes 24h post-vaccination

We and others previously described how AS03 expands antibody response diversity by eliciting antibodies to influenza clades beyond those included in the vaccine[213,214 282]. Further analysis of post-vaccination serum revealed antibody binding avidity to HA protein of the influenza virus strain in the vaccine was highly correlated across individuals with binding to a non-vaccine strain HA protein (Fig. 3.5i), an effect only seed in individuals vaccinated with AS03. This suggested expansion of B cell clones did not emerge at the expense of strain-specific immunity, but rather that AS03 may have nonspecifically reduced B cell selection constraints. Suppression of apoptosis in naïve B cells was also specific to subjects vaccinated with AS03, including of a downregulation of a lymphocyte turnover module (Fig. 3.6b) and of canonical apoptosis gene NOXA (PMAIP1), which was upregulated in non-adjuvanted subjects (Fig. 3.5 j). NOXA deficiency is known to increase lymphocyte repertoire diversity[304,305], for example B cells from NOXA$^{-/-}$ mice outcompete wild type cells for entry into the germinal center following influenza vaccination and infection, these cells subsequently persist due to inefficient apoptosis[305] and increase diversity of anti influenza antibodies. The naïve B cells in humans after vaccination with AS03 may thus phenocopy those of influenza vaccinated NOXA$^{-/-}$ mice through a similar process. Supporting this hypothesis, single cell analysis indicated naïve B cells from subjects vaccinated with AS03 had increased expression of a gene signature derived from CD40 activated B cells[306,307] with the opposite direction of change following vaccination in unadjuvanted subjects on day 1 (Fig 3.6c). Both apoptosis and CD40 activation had conserved direction change within

sorted total B cells in the validation cohort comparing AS03 to the PBS control (Fig 3.6d), though apoptosis was not significant after FDR correction owing to the original CITE-seq signal being derived from changes in a small fraction of only naïve B cells. The day 1 fold change in CD40 activation in naïve B cells across individuals was also negatively correlated with the apoptosis signature (Fig. 3.5k), suggesting phenotypic coupling of these states. Together these results highlight two potential mechanisms by which AS03 may drive desirable antibody responses. Heightened innate pathogen sensing capacity could drive recruitment and activation of cells presenting antigen and interacting with B cells, which in turn suppress apoptosis potentially initiating more naïve B cells into the germinal center reaction to increase antibody breadth.

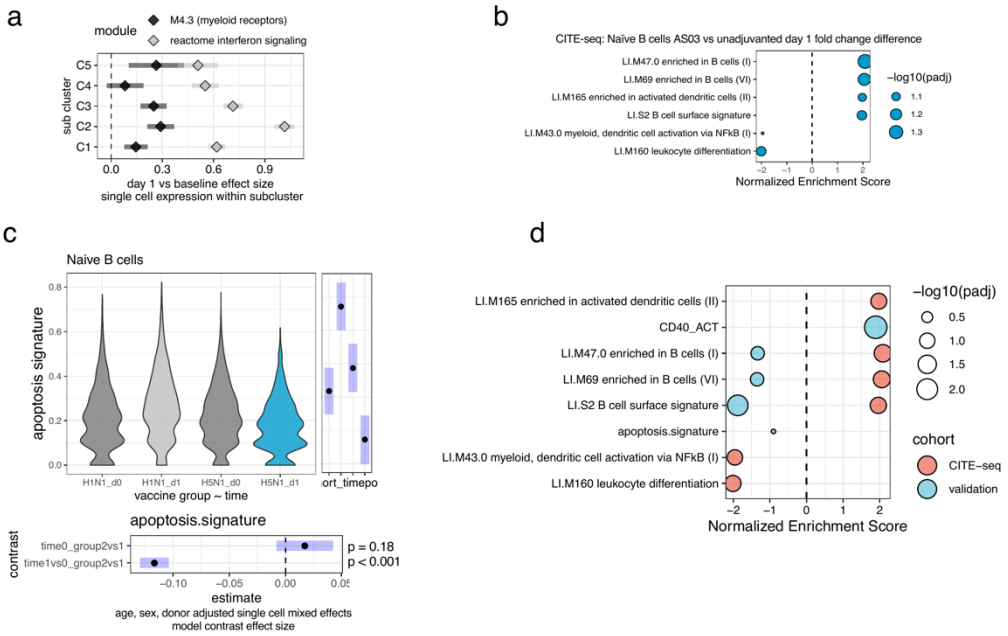


Figure 3.6 External cohort validation of AS03 perturbation phenotypes and additional analysis AS03 induced lymphocyte phenotypes

a. Mixed effects model of module-level fold change of single cell gene module scores for AS03 specific leading edge genes tested in monocyte sub-clusters, positive effect size indicates upregulation across individuals; individual modeled as a random effect. **b.** As in Fig 3.5b for Naïve B cells. **c.** Single cell mixed effects model of differential expression between groups as a function of time post vaccination. The effect size for the time effect for each cohort was opposite, (bottom contrast on bottom margin of plot). The right margin shows the estimated marginal means of the model averaged over levels of covariates. **d.** Naïve B cell derived perturbation phenotypes (leading edge genes from CITE-seq based enrichments of contrast model) tested in validation cohort of total sorted B cells (Naïve B cell AS03 specific pathways from CITE-seq data tested vs all CD19+ cells in the validation cohort).

3.10 A broadly activated multi cell type coupled immune cell network defines the baseline immune set point of high antibody responders

We previously described overlapping baseline immune set points between SLE patients and healthy subjects[115]. Since that analysis was constrained to dissect the biology of bulk predictive signature linked to B cell phenotypes, here we devised an unbiased comparative analysis of baseline immune cell phenotypes associated with increased antibody responses. We first contrasted transcriptome phenotypes robustly enhanced in high vs low responders, then devised a network analysis approach to define their inter and intracellular coupling across subjects based on shared information in gene expression (see methods). High responder effector lymphocyte and innate cell phenotypes could be grouped into 9 functional categories which together defined a broadly activated high responder immune setpoint network (Fig. 3.7a, Fig. 3.8a). Two highly connected cell phenotypes network are highlighted in (Fig. 3.7 b, c). Within CD14 monocytes, the "FC receptors and phagocytosis" pathway included specific genes involved in sensory capacity (e.g. FCGR3A, FCGR1A, FCGR2A), regulators of cytoskeletal reorganization active during phagocytosis (e.g. PAK1, ARPC5 CFL1 ARF6) and genes reflecting activation related to second messenger signaling (PIP5K1A, PIK3CD AKT1, MAPK12, ARPC2). This monocyte phenotype elevated in high was coupled to 27 other cell phenotypes elevated in high responders (adjusted $p < 0.05$), and was thus a high degree node "hub" in the setpoint network (Fig. 3.7b, (Fig. 3.8b). The distribution of expression across high (red) and low responders (blue) is shown for two network edges, which reflect how the CD14 monocyte FC receptor state was correlated with both antigen presentation in naïve B cells and the interferon response in CD16 monocytes (Fig. 3.7b, bottom). This CD16 monocyte interferon response phenotype itself was hub in the network, coupled to 28 other cell phenotypes, including those within a CD8-CD4-CD161+ T cell cluster, phenotypically similar to MAIT cells (Fig. 3.7c). CD161 high T cell subsets are known to have tissue homing capacity[308] and MAIT cells can also act as sensors of bacteria through their invariant TCR and TLRs[254,309] yet also limit lethal influenza infection *in vivo*[310] in a manner dependent on interferon producing cytokines IL-18 and IL-12[311]. The interferon response pathway in CD16 monocytes was correlated with interferon response and cytokine signaling pathway in these MAIT-like cells (Fig. 3.7c, bottom) which reflected shared baseline upregulation of IFITM1, IFITM2, ISG15 and IFI6. Furthermore, the

quantitative level of the baseline states including cytokine signaling in MAIT cells and the FC receptor phenotype of CD14 Monocytes was tightly coupled to the quantitative level of the day 7 plasmablast activity signature in blood (Fig. 3.7d). These high degree nodes thus capture both correlated activity of the setpoint in different cell types across individuals, and correlated activity of later antibody producing plasmablast cell responses on day 7 which itself is predictive of day 70 antibody levels.

How baseline set point phenotypes are themselves kinetically altered within the same immune subsets by vaccination is unknown, limiting our understanding their putative functional relevance. Given the similarities between phenotypes comprising this baseline innate setpoint network of high responders and the states coherently induced within subsets following vaccination (Fig. 3.4a) we directly modeled the early (24h) post–vaccination kinetics of the baseline genes elevated in high responders in a cell type specific manner. Using a single cell mixed effects model adjusting for several covariates (see methods) we found that the same genes comprising the baseline high responder setpoint network in CD14 and CD16 monocytes, mDCs and MAIT cells were themselves coherently induced across individuals by vaccination within the same subsets (Fig. 3.7e). This suggested that the innate network setpoint reflecting a naturally activated immune state may have primed responses to vaccination since it was itself induced by vaccination.

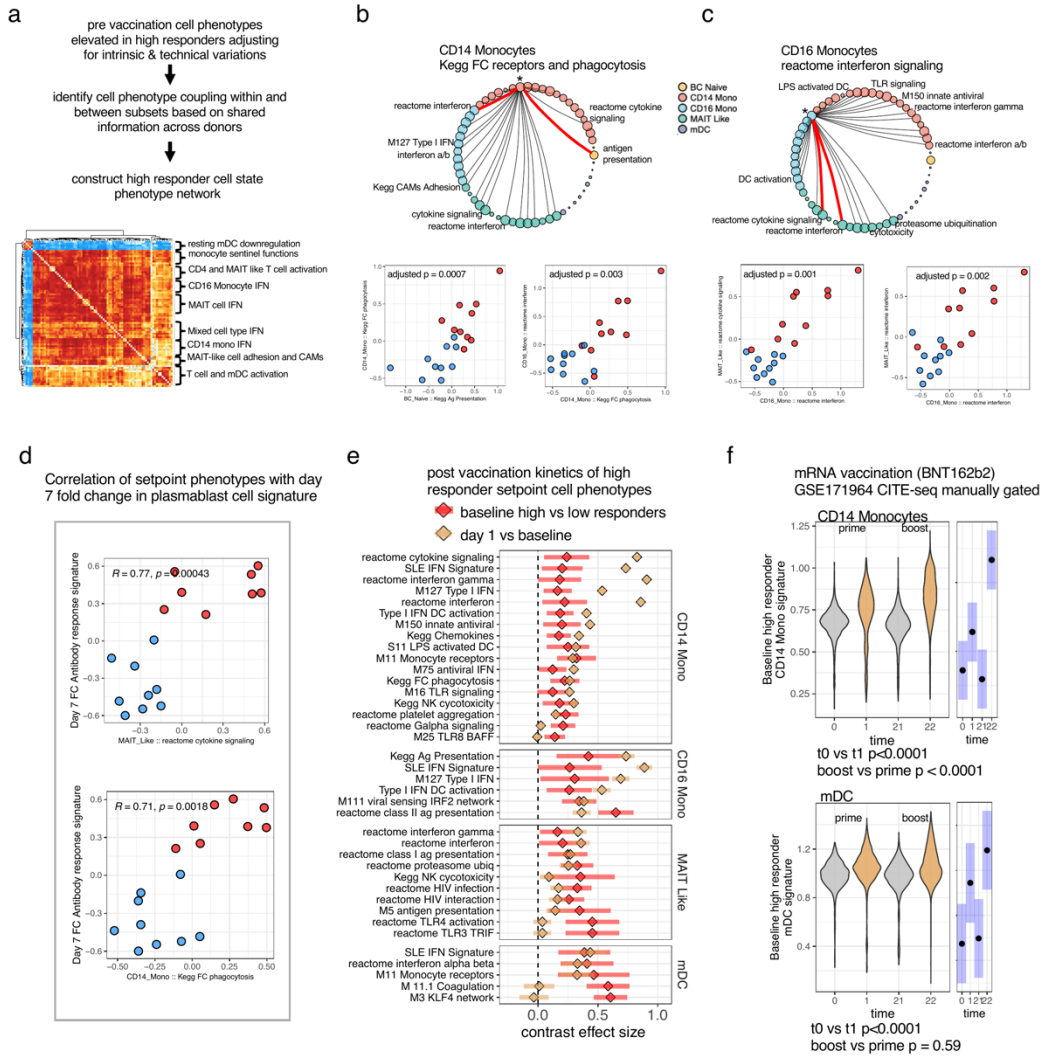


Figure 3.7 The immune setpoint network phenotypes of high responders their day 1 post-vaccination kinetics and correlation with plasmablast activity

a. Construction of the baseline high responder setpoint network. First gene set enrichment analysis of modules enriched pre-vaccination (baseline) in high vs. low responders within each cell type, adjusting for age, sex, batch. The leading edge genes from these enrichments were correlated across donors within and between cell types. Within a cell type, the Jaccard similarity of each pairwise leading edge gene was subtracted from the spearman correlation coefficient to correct for correlation due to two signals sharing the same genes (within a cell type) and high confidence high connectivity edges were retained in the network (see methods). **b-c** Two selected highly coupled cell phenotypes in the high responder setpoint network. The edges highlighted in red are shown below as correlations of the activity of the leading edge genes from those modules across donors within the cell type indicated by the edge. **d.** The correlation of signature expression within cell types with the day 7 fold change in the predictive signature we previously found was predictive of antibody response associated with plasmablast activity from microarray data. **e.** The post vaccination kinetics of the components of the high responder innate setpoint network. A single cell mixed effects model of module activity was used to estimate the baseline high vs low responder effect size (red) and day 1 fold change across subjects adjusting for age, sex, number of cells per donor and modeling individual with a random effect. **f.** Day 1 vs 0 prime and day 22

vs 21 boost kinetics of baseline high responder states tested in an external cohort of monocytes and DCs manually gated from CITEseq data (GSE171964) collected on individuals vaccinated with mRNA vaccine BNT162b2. The difference in the fold change between boost (d22 vs d21) and prime (d1 vs d0) p values: mDC 0.59, CD14 monocyte < 0.001.

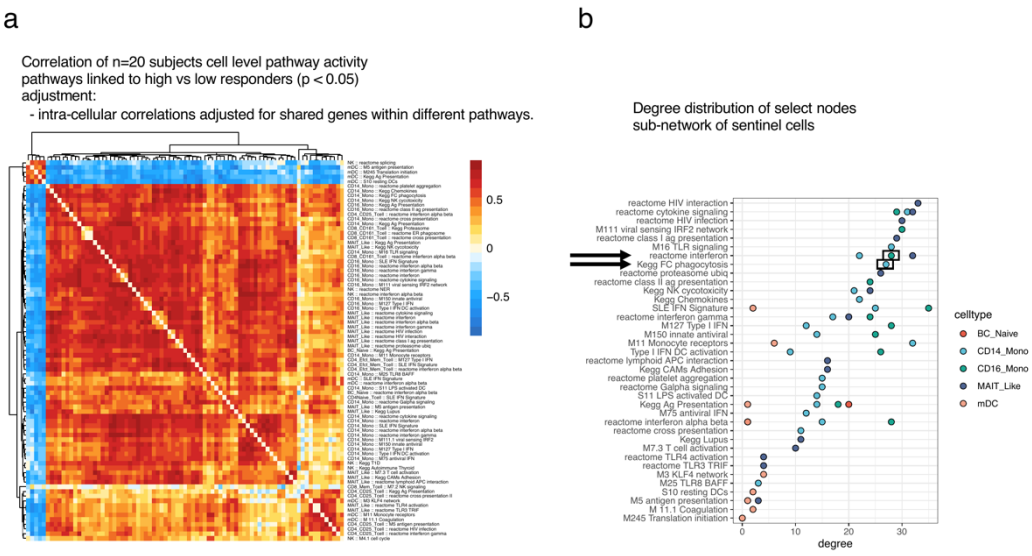


Figure 3.8 Additional information on the immune setpoint network of high responders

a. Spearman correlation matrix of cell phenotypes associated with high vs low responders as shown in Fig 3.7a with more detail. **b.** The degree distribution of nodes in the high responder setpoint network. The nodes highlighted in (Fig 4 b-c are indicated with a black arrow and box. Points are colored by cell type; the annotation of modules may be the same for a given row (e.g. reactome interferon in CD14 and CD16 monocytes) but the same module is captured by different genes driving the high responder effect in each cell type (e.g. they reflect cell type specific cell phenotypes).

We next investigated whether the setpoint network was also induced in monocytes and mDCs following mRNA vaccination, which emerged as efficacious formulations during the covid19 pandemic[312]. Reanalysis of CITE-seq data collected from six individuals[313] vaccinated with the BNT162b2 mRNA SARS-Cov2 vaccine demonstrated setpoint genes of CD14 monocytes and mDCs were also induced day 1 across n=5 individuals within protein gated mDC and CD14+ monocytes after mRNA vaccination. Furthermore, the fold change of the monocyte setpoint genes were elevated to a greater extent after the second dose (day 21 vs 22) than after dose 1 of mRNA vaccination (Fig.

3.7f) in classical monocytes. This further demonstrated the baseline setpoint circuitries could themselves reflect a naturally primed state which enhances immune response potential. The mRNA vaccine lipid nanoparticle (LNP) carrier is thought to act an adjuvant[312] given that protein vaccines formulated with LNPs increase GC B cells and HAI titers[314] and when administered without mRNA, LNPs alone recruit both classical monocytes and DCs to the injection site. Interestingly these same phenotypes we identified as being perturbed by AS03 and primed at baseline in high responders were also the only two cell types which translate the mRNA vaccine antigen[315]. Priming of the high responder setpoint phenotypes in particular within classical monocytes by mRNA vaccination suggests high responder innate network circuitries acts like a "natural adjuvant", increasing innate immune potential prior to stimulation.

3.11 High responders have a naturally adjuvanted baseline immune system set point

We next formally tested whether high responder DCs and monocytes were also primed with processes specifically induced by the AS03 adjuvant (as an experimental model of adjuvantation). As a group, these AS03 specific phenotypes defined as genes driving pathway enrichments from the AS03 specific model (shown in Fig. 3.5b) were in fact downregulated after unadjuvanted vaccination, suggesting they reflected distinct states from the unadjuvanted response kinetics observed above (Fig. 3.9a). Further pruning genes with validated adjuvant specificity from analysis of the validation cohort comparing AS03 to PBS demonstrated conserved enrichment including of sentinel genes highlighted in Fig 3.5c-d. Consistent with the naturally adjuvanted hypothesis, AS03-specific cell phenotypes were elevated at baseline in high responders compared to low responders in both mDCs and CD14 monocytes. A separate study of AS03 identified increased circulating frequencies of activated HLADR+ cells including monocytes 24h following vaccination[159]. Having established that the internal monocyte transcriptional phenotype reflected baseline elevation of states coherently induced by vaccination, we examined the baseline surface activation status of unadjuvanted donors monocytes at baseline and found that again high responders appeared to phenocopy AS03 cellular responses, with increased HLADR+ monocyte frequency at baseline detected using flow cytometry data from our full cohort[39] (Fig. 3.9d). Furthermore, post-vaccination HLADR kinetics revealed high responders elevated post vaccination

activated HLADR+ monocyte frequency to a greater extent than low responders especially at day 1 (Fig. 3.9e), suggesting amplification of immune state biased toward stronger responses. This further suggested high responder setpoint circuitry reflected naturally adjuvanted innate immune cells.

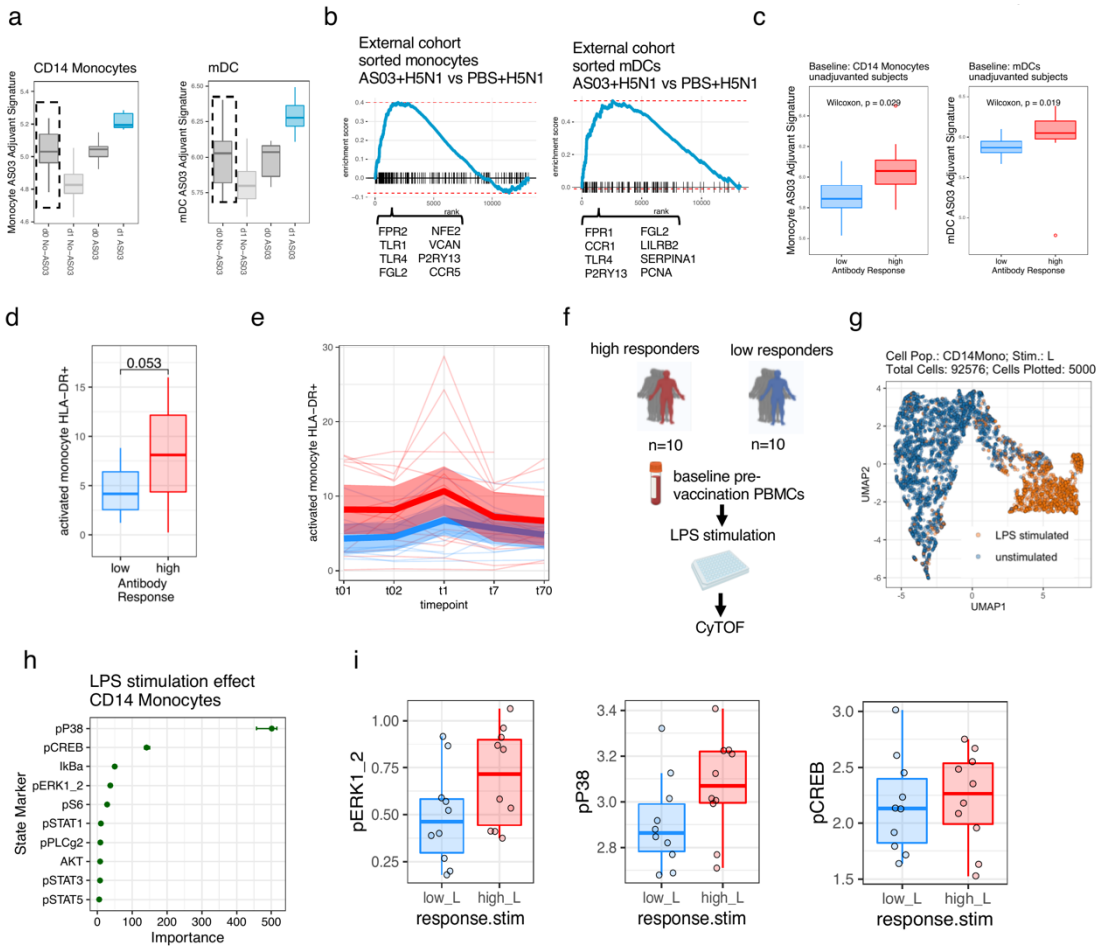


Figure 3.9 High responders have a naturally adjuvanted immune competent setpoint with cells more poised to respond to innate stimulation

a. Average expression of a combined gene signature reflecting the AS03 specific induced states within DCs and CD14 monocytes. **b.** Gene set enrichment of the combined AS03 specific signature on the validation cohort in analogous subsets; select sentinel genes in the leading edge of the validation are shown. **c.** The average expression in high vs low responders of the mDC and CD14 monocyte AS03 specific day 1 induced validated signature tested in analogous subsets. **d.** Log cell frequency of HLA-DR+ classical monocytes as a percentage of total classical monocytes in high vs low responders at baseline, p value from a Wilcoxon rank test. **e.** The kinetics over two baseline timepoints and three post vaccination timepoints for HLA-DR+ classical monocytes. Mixed effects model with an interaction for time and response group and a random effect for subject ID–high responder effect size 3.17 p value = 0.0005, low responder effect size 1.89, p value = 0.14, difference in estimated marginal day 1 vs baseline fold change not significant, response time vs time only interaction model ANOVA p = 0.063. **f.** Schematic outlining CyTOF stimulation experiment. PBMCs isolated from high and low responders were stimulated with PRR ligands. Stimulation

phenotype and markers driving stimulation were defined with HDStIM. **g.** UMAP plot of a random subset of 5000 monocytes pre and post stimulation with stimulated cells in orange (unstimulated = blue). **h.** Variable importance for automatic determination of responding cells from HDStIM. **i.** The post stimulation median marker intensity of phosphor markers within the CD14 monocyte cluster, the post stimulation aggregated data are shown, effects tested using a mixed model adjusting for batch and modeling individual with a random effect. Difference in pre vs post stimulation fold changes in high vs low responders contrast estimate and p values: p38 contrast effect: 0.104, p = 0.058, pCREB contrast effect: 0.223, p = 0.024, pERK 0contrast effect: 0.58, p = 0.055. Note only the post stimulation timepoint is shown.

3.12 The naturally adjuvanted setpoint is an immunocompetent state with sentinel cells more able to kinetically signal PRR ligands

Transcriptional circuitries forming the high responder naturally adjuvanted setpoint included correlated sensory capacity and primed IFN response potential in innate cells, both linked to later plasmablast expansion and antibody response (Fig. 3.7a-d), and elevation of adjuvant induced phenotypes which also in part reflected innate cell sensory function. To further understand naturally occurring variation in the earliest innate cell protein signaling kinetics in response to antigenic ligands sensed by these cells, we *ex vivo* stimulated PBMCs from the same 10 high and 10 low responders with multiple innate PRR ligands (Fig. 3.9f). We used Cy-TOF profiling for cell surface and intracellular phosphor-signaling readouts, and defined responding cell and markers driving stimulation phenotypes automatically using a computational algorithm, HDStIM[316] (Fig. 3.9g-h). CD14 monocytes were strongly perturbed by LPS stimulation, driven by phosphorylated p38, CREB, IkBa, and ERK (Fig. 3.9h). Supporting the idea that the naturally adjuvanted setpoint also reflected cell intrinsic signaling capacity, the difference in the post-stimulation fold change of p38, pERK and pCREB to LPS, adjusting for batch and including a donor random effect, was higher in high vs low responders (Fig. 3.9i). Together these results provide further insights into how naturally adjuvanted individuals innate cells are primed to improve antibody generation following vaccination. The cell intrinsic phosphoprotein signaling capacity of monocytes may have acted to both establish the observed transcriptional setpoint as well as amplify early post vaccination innate responses. For example, the main ligand of LPS (TLR4) signals through IRF3 to activate interferon response genes including ISG15 and IFN-b which activate antiviral gene expression programs in an autocrine / paracrine fashion[317]. Elevated cell intrinsic signaling potentially enhanced through signals sent through elevated basal innate cell pathogen sensors (Fig. 3.7a-b) could have primed pre-

vaccination IFN response programs (Fig. 3.7a-c) which fed forward to elevate baseline antigen presentation capacity (Fig. 3.7a,b, Fig 3.9e). These basal interferon response processes were further triggered by vaccination (Fig 3.7e-f) and likely further amplified by inflammatory effectors (Figs. 3.4 a-c, f-g) leading to enhanced monocyte antigen presentation capacity 24h later (Fig. 3.9d,e) plasmablast expansion on day 7 (Fig. 3.7d) and ultimately, elevated antibody response.

3.13 Discussion of Chapter 3 results

Biomarker signatures derived from immune system profiling studies can form the basis of clinical stratification tools[80], or hypothesis generation for further targeted study. In this work, we demonstrated a framework for using CITE-seq to integrate information across biological scales, down to the level of single cells to improve insight and hypotheses gleaned from top-down profiling. Our approach identified multicellular transcriptome circuitries of high responders poised toward baseline activation of early response genes induced by vaccination, including processes which were specifically induced by adjuvant. Together, these processes formed the basis of a naturally adjuvanted baseline immune setpoint, analogous to a compressed spring, with cells poised to respond to innate stimulation. The discovery of the naturally adjuvanted state provides detailed molecular circuits as immune engineering targets which advances the concept we proposed of altering immune setpoint circuitries toward those which improve response outcomes in a variety of contexts[278]. Furthermore, our approach here provides a proof of principle for utilizing human population variation to define precision medicine targets by applying these multiscale models to link molecular features to emergent response outcomes, a long standing goal of human systems immunology approaches.

Our study has several limitations. Profiling tissues such as lymph nodes could give a more complete picture of vaccination response variations. Despite the logistical challenges of profiling human tissues, recent pioneering work collected lymph nodes following influenza vaccination, revealing germinal center B cells were a mix of naïve and memory in origin[318]. While that proof of concept study was small (n=8) and single cell profiling was run on only one individual, the computational framework developed

for this study could be applied to future work integrating blood and tissue data from hundreds of individuals to contrast vaccine formulations on molecular phenotypes. With respect to linking the phenotypes from blood to tissue, our analysis revealed predictive day 7 bulk expression signatures were derived nearly exclusively from a small number of plasmablast cells; the lymph node study by Ellebedy and colleagues[318] revealed these same circulating plasmablast clones do indeed transit to the lymph node as expected. In that study, B cell clones were computationally tracked through shared recombined lymphocyte receptor sequences[319]. A natural extension applied to findings of our work here include determining the lineage origin of the innate cells in circulation on day 1, including the sensory/sentinel cell states induced by AS03 and the naturally adjuvanted phenotypes forming the high responder setpoint. Tracking the precise clonal origins of innate cells lacking recombined receptors in humans presents a major challenge, though recent developments in mitochondrial gene profiling of single cell ATAC-seq data could be informative in this context[320].

Baseline immune biomarkers have been increasingly linked to outcomes[115,223,277,278] however, little is known about the cellular origins or dynamics of these cellular processes following an immune trigger. Here we pinpointed the specific antibody response associated innate cell set point circuitries and found they were themselves induced by vaccination, including being enhanced further after initial priming by mRNA vaccination. This suggests either an ongoing or a previous immune exposure could be responsible for tuning the naturally adjuvanted setpoint. We recently participated in a consortium meta-analysis of bulk transcriptome data which identified baseline signatures linked to antibody responses overlapping with bulk gene expression signatures seen during bacterial sepsis likely stemming from innate cells as well[225]. This suggests tonic signals derived from the microbiome[321] or past infection could tune innate cell tone, interestingly, bacterial depleting antibiotic treatment in humans only impacted antibody responses to a strain of influenza in naïve individuals[180]. Using epigenome profiling, it was also recently shown that vaccination with AS03 induced alterations in chromatin accessibility[322] which overlapped alterations linked to potentiation of innate immune memory[174,176], and the IRF family. Using an “orthogonal omic integration” approach[323], we predicted transcription factors controlling the genes in AS03 specific DC and classical monocyte signatures which overlapped with the

naturally adjuvanted high responder setpoint. Predicted transcription factors included PU.1 and members of the CEBP family[175] (data not shown) which have been linked to potentiation of innate responses following prior exposure. Human studies comparing engineered setpoints induced by different vaccines are needed to further resolve these interesting chromatin and transcriptome circuitries linked to improved secondary responses. Together, our dataset, software pipeline, and curated findings help advance vaccine studies toward a quantitative, predictive understanding of human immunology, and pave the way for further in vivo single cell drug / perturbation screens across biological disciplines.

3.14 Methods – Chapter 3

Human vaccination comparison cohorts and antibody response assessment

Healthy volunteers were enrolled on the National Institutes of Health (NIH) protocols 09-H-0239 (Clinicaltrials.gov: NCT01191853) and 12-H-0103 (www.clinicaltrials.gov: NCT01578317). Subjects enrolled in 09-H-0239 received the 2009 seasonal influenza vaccine (Novartis), and the 2009 H1N1 pandemic (Sanofi-Aventis) vaccines, both without an adjuvant. Subjects in 12-H-0103 received a vaccine formulated with the adjuvant AS03 containing avian influenza strain H5N1 A/Indonesia/05/2005 (GSK). In both cohorts, virus neutralizing antibody titers assessed using a microneutralization assay were determined as previously reported. The highest titer that suppressed virus replication was determined for each strain in the 2009 inactivated influenza vaccine: A/California/07/2009 [H1N1pdm09], H1N1 A/Brisbane/59/07, H3N2 A/Uruguay/716/07, and B/Brisbane/60/2001 or for AS03 adjuvanted influenza vaccine, H5N1 A/Indonesia, clade 2.1. High and low antibody responders to the unadjuvanted vaccination were defined using the adjusted maximum fold change (AdjMFC) which adjusts the fold change for the baseline antibody titer (methodological details in the supplementary methods of our previous report[39]). In the unadjuvanted cohort, n=10 high responders and n=10 low responders were selected for CITE-seq profiling. All subjects were analyzed pre–vaccination, with a subset of 8 and 12 donors profiled on days 1 and 7 post-vaccination also split evenly between high and low responders. In the adjuvant cohort, n=6 subjects with robust titer responses were selected for CITE-seq.

CITE-seq profiling peripheral blood mononuclear cells

We optimized a custom CITE-seq antibody panel of 87 markers using titration experiments and stained cells with a concentration of antibody appeared to saturate ligand of the cell population with the highest marker expression, or used the manufacturers recommended concentration when below saturation. We stained the 52 PBMC samples across three experimental batches using a single pool of which were combined in the optimal concentration and concentrated in an Amicon Ultra 0.5mL centrifugal filter by spinning at 14,000 x g for 5 minutes. Three aliquots of 12µL from the 36µL volume of optimized antibody mixture was used on 3 subsequent days to minimize between experiment technical variability. Frozen PBMC vials from each donor were washed in pre-warmed RPMI with 10% FBS followed by PBS. $1x10^6$ cells from each sample were stained with a hashing antibody[98] simultaneously with 1µL FC receptor blocking reagent for 10 minutes on ice. After washing the hashing reaction 3 times in cold PBS, cells were counted and pooled in equal ratios into a single tube and mixed. The sample pool was concentrated to $5x10^6$ cells in 88µL of staining buffer. 12µL of the concentrated optimized 87 antibody panel was added to stain cells (total reaction volume 100µL) for 30 mins on ice. After washing cells, we diluted cells to 1400 cells / µL, recounted 4 aliquots of cells and 30µL of the stained barcoded cell pool containing cells from all donors was partitioned across 6 lanes of the 10X Genomics Chromium Controller for each of the 3 batches for 18 total lanes. We proceeded with library prep for the 10X Genomics Chromium V2 chemistry according to the manufacturer's specifications with additional steps to recover ADT and HTO libraries during SPRI bead purification as outlined in the publicly available CITE-seq protocol https://cite-seq.com (version 2018-02-12). We clustered Illumina HiSeq 2500 flow cells with V4 reagents with pooled RNA, ADT and HTO libraries in a 40:9:1 ratio (20µL RNA, 4.5 µL ADT, 0.5µL HTO). Libraries were sequenced using the Illumina HiSeq 2500 with v4 reagents.

CITE-seq data bioinformatic alignment and sample demultiplexing

Bcl2fastq version 2.20 (Illumina) was used to demultiplex sequencing data. Cell Ranger version 3.0.1 (10x Genomics) was used for alignment (using the Hg19 annotation file provided by 10x Genomics) and counting UMIs. The fraction of reads mapped to the genome was above 90% for all lanes and sequencing saturation was typically around

90%. ADT and HTO alignment and UMI counting was done using CITE-seq-Count version 1.4.2. We retained the "raw" output file from Cell Ranger containing all possible 10X cell barcodes for each 10X lane, and merged the CITE-seq-count output. For each 10X lane, barcodes were concatenated with a string denoting the lane of origin and data for ADT, HTO and mRNA. We then utilized combined sample demultiplexing to assign the donor ID and timepoint to each single cell. Both the timepoint and response class were identifiable based on the hashing antibody. The first round of demultiplexing was carried out via cell hashing antibodies. The union of singlets defined by the multiseq deMUTIplex procedure[324] and Seurat's HTODemux function were retained for further QC. Negative drops identified by HTODemux were retained for further QC and use in denoising and normalizing protein data. The second round of sample demultiplexing was carried out via Demuxlet[99] to assign the unique donor ID by cross-referencing unique SNPs detected in mRNA single cell data against a vcf file with non-imputed illumina chip based genotype data from the same donors. Demuxlet provided an additional round of doublet removal via an orthogonal assay (mRNA) to antibody barcode (HTO) based demultiplexing thus providing further data QC. Only cells that met the following conditions were retained for further downstream QC, normalization and analysis: 1) The cell must be defined as a "singlet" by antibody barcode based demultiplexing and by demuxlet. 2) The identified donor from demuxlet must match one of the expected donors based on cell hashing. Cells were then further QCd based on mRNA using calculateQCmetrics function in scater[325]. Cells were removed that had with greater or less than 3.5 median absolute deviations from the median log mRNA library size.

mRNA and surface Protein count data normalization

We denoised and normalized ADT data using an open source R package we developed for this work called dsb[100] which removes noise derived from ambient unbound antibodies and cell to cell technical noise. We used function DSBNormalizeProtein with default parameters. We normalized mRNA on the entire dataset with the normalizeSCE and multiBatchNorm functions from scran[326] using library size-based size factors. Various analysis utilized aggregated mRNA data which was separately normalized for analysis at the subset level as a "pseudobulk" library; single cell mRNA data were also rescaled for specific analysis as outlined below.

Protein-based clustering and cell type annotation

Using protein to define cell type facilitated improved interpretation of transcriptome differences between vaccination groups. Cell types were defined with statistically independent information, protein, from transcriptome data being modeled within each cell type (Fig. 1a). We clustered cells directly on a distance matrix using the parallelDist package calculated from the non-isotype-control proteins all cells using Seurat's FindClusters function using parameters: res =1.2, modularity.fxn = 1, algorithm = 3 (SLM[273]). We annotated cell types in the resulting clusters post hoc based canonical protein expression in immune cell populations. This procedure improved separation of known immune populations compared to compressing protein data using principal components as commonly done for higher dimensional mRNA data (data not shown). Analysis of unadjuvanted vaccination responses was first done blind to the adjuvanted cohort data. We thus first applied high dimensional clustering of the unadjuvanted cohort and annotated cell types with additional manual gates to purify canonical cell populations such as memory and naïve T cells. We next merged unadjuvanted and adjuvanted cohort cells and used annotations to guide combined clustering annotation, again manually refining cell populations using biaxial gating scripts in R to purify cell some cell populations. For annotation, the distribution of marker expression within and between clusters was compared using density histogram distributions of marker expression across clusters at the single cell level, biaxial marker distribution and median and mean aggregated protein expression across clusters.

Hierarchical transcriptome variance deconstruction to intrinsic individual, cell type and vaccine effects

To estimate the contribution of intrinsic and experimental factors to the total variation in expression of each gene, we used the variancePartition package. The set of models used for estimating variance fractions are distinct but related to those used for testing differential expression and contrast vaccination effects within subsets (see below). We first aggregated data by summing expression for each individual, timepoint and cell type. The normalized aggregated expression was used to model the mean variance relationship using observation level weights using voom. Mixed effects linear models of the expression of each gene across the aggregated libraries were then fitted using lme4.

For example, for a given gene “y” the total variance was defined by 780 measurements derived from the 58 PBMC deconvolved into 15 protein-based cell types tested. The model fit to each gene “g” was:

$$g = \sum_j X_j \beta_j + \sum_k Z_k a_k + \varepsilon_g$$

Where X and Z are the matrices of fixed and random effects respectively, and random effects are modeled with a Gaussian distribution and errors incorporate voom weights.

$$a_k \sim N(0, \sigma_\alpha^2)$$

$$\varepsilon_g \sim N\left(0, diag\left(w_g\right)\sigma_\varepsilon^2\right)$$

The variancePartition package then incorporates both fixed and random effects in calculating the fraction of variation attributable to each variable in the model. For example, the variance in g attributable “subjectID” denoting the subject of origin which was modeled as a random effect is equal to:

$$\sigma_{gSubjectID}^2 = \frac{\sigma_{\beta_{SubjectID}}^2}{\sum_j \sigma_{\beta_j}^2 + \sum_k \sigma_{\alpha_k}^2 + \sigma_e^2}$$

The denominator in the fraction above is the total variance of gene g, and each variable in the model contributes a fraction of the total variance which together always sum to 1. In the first model above, age sex, subjectID, timepoint, response, and a cell type and timepoint interaction term were included with categorical variables as random effects as required by the variancePartition framework. A second set of models fit within each cell type increased the apparent variance explained by the experimental factors independent of major cell type specific expression driving gene variation. This model included age sex, subjectID, timepoint, and response/vaccine group (unadjuvanted group high vs low responders, or AS03 group) and an interaction term for time and group.

Within cell type linear mixed effect models of vaccination effects on gene expression

We next used linear mixed models to test vaccination effects while adjusting for intrinsic and individual level variation. Gene expression counts were aggregated within

each protein based cell type by summing counts for each sample. The lowest frequency cell types without representation across individuals and time relative to vaccination (e.g., HSCs, donor-specific cell types, or plasmablasts which were mainly detected on day 7) were excluded from this analysis. Three main analysis were carried out to model gene expression within each cell type to estimate the following vaccination effects over time across individuals: 1) unadjuvanted subjects day 1 vs baseline, 2) unadjuvanted subjects day 7 vs baseline, 3) A contrast of the difference in day 1 fold change between unadjuvanted and adjuvanted subjects in a combined model. All models were fit with the 'dream' method[71] which incorporates precision weights[73] in a mixed effects linear model fit using lme4[72]. The models included a random intercept for subject ID to adjust for variation in baseline expression and non-independence of repeated measures from the same individuals. Models For models 1 and 2 above (unadjuvanted vaccination effects) we fit the following model: *gene ~ 0 + time + age + sex + (1|subjectID)*.

The fitted value for expression y of each gene g corresponds to:

$$y_g = \beta_{0g} + \sum_j X_j \beta_j + \varepsilon_g$$

With variables time, age and sex represented by covariate matrix X. The β_0 term corresponds to the varying intercept for each donor represented by the (1|subjectID) term. This models variation across subject baseline expression S_0 around the average γ_0 using a Gaussian distribution with standard deviation ${\tau_g}^2$. Errors ε_g incorporate precision weights w_g calculated using the function *voomWithDreamWeights* as below:

$$\beta_{0g} = \gamma_0 + S_0$$

$$S_0 \sim N\left(0, {\tau_g}^2\right)$$

$$\varepsilon_g \sim N\left(0, diag\left(w_g\right)\sigma_\varepsilon^2\right)$$

In this model the day 1 or day 7 effect across subjects was the time effect from the model. The effect size was then used to rank genes for enrichment testing for each cell type.

Model 3 was specified as *gene ~ 0 + group + age + sex + (1|subjectID)*. The “group” variable corresponds to a combined factor representing the vaccine formulation received (adjuvanted vs unadjuvanted) and timepoint (baseline or day 1 post vaccination). with 4

level: “d0_AS03”, “d1_AS03”, “d0_unadjuvanted”, “d1_unadjuvanted”. A contrast matrix L_{delta} corresponding to the difference in fold changes between adjuvanted and unadjuvanted subjects was applied to the fitted values:

$$L_{delta} = [\ -1 \quad 1 \quad 1 \quad -1 \quad 0 \quad 0\ ]$$

With the first four columns representing the group factor and the two 0s representing age and sex effects. The contrast fit outputs the difference in fold change after adjusting estimates for age, sex and donor variation with positive effects representing increased fold change in the adjuvant group compared to the unadjuvanted group. This also captures genes with opposite effects in the two groups, for example, upregulation in the AS03 group and downregulation in the nonadjuvanted subjects.

Transcriptome data was uniformly processed for all fitted models above. Aggregated (summed) single cell UMI counts were normalized within each protein based cell type using the trimmed means of M values method. Genes were retained which had a pooled count per million above 3 using the edgeR *filterByExprs* function; cell type specific gene filtering removed genes non expressed by each lineage from analysis ensured the model assumptions used to derive precision weights and account for the mean variance trend were met. We verified the log count per million vs. voom fitted residual square root standard deviation had a monotonically decreasing trend within each cell type. Models were fit with wrappers around functions, *dream*, and *eBayes* from the variancePartition package. For the AS03 validation cohort, pre normalized data were downloaded from the study supplemental data and a similar model to model 3, contrasting the difference in fold change was fit with a contrast again using a donor random intercept.

Enrichment testing of vaccination effects within cell types using hypothesis sets and unbiased pathways

To test enrichment of pathways based on the estimated gene coefficients corresponding to the three vaccination effects defined above, we performed gene set enrichment analysis using the fgsea[327] package splitting monte carlo algorithm. Genes for each coefficient (i.e. models 1-3) and each cell type were ranked by their effect size, (the empirical Bayes moderated signed z statistic), corresponding to pre vs post vaccination

or the difference in fold change for model 3. For day 1 enrichment, a set of 5 signatures derived from bulk transcriptome profiling influenza vaccination (see supplementary table), and an additional 25 pathways curated from public databases were tested. For Day 7 and the difference in fold change between adjuvanted and unadjuvanted subjects, an unbiased set of pathways were tested from Li et al. Blood Transcriptional modules[88], MSigDB Hallmark, reactome and Kegg databases. Individual gene subsets, for example category and pattern genes not based on a full ranked list were tested for enrichment using enrichr[190].

Derivation of the high responder baseline immune setpoint network

To define cell phenotypes robustly associated with high vs low responders to the unadjuvanted vaccine at baseline, we used limma[70] to fit a linear model of antibody response (high vs low) adjusting for age sex and batch (e.g. in R symbolic notation, gene ~ AdjMFC + age + sex + batch) as fixed effects on aggregated (summed) data for each cell type, similar to models above without varying effects for subject ID, i.e.:

$$y_g = \sum_j X_j \beta_j + \varepsilon_g$$

Errors incorporated voom weights as above. Gene coefficients for each cell type corresponding to model adjusted empirical Bayes regularized estimates for high vs low responder effect at baseline were input into gene set enrichment analysis against the unbiased set of pathways described above. We then calculated the average module z score[115] using log counts per million from each cell type of the high responder associated cell phenotypes (using only high responder associated leading edge genes), resulting in a matrix of baseline normalized expression of pathways across 20 individuals (10 high and low responders) for each cell type specific signal. We next tested for phenotypic coupling of these signals within and between cell types by calculating the spearman correlation and correcting p values with the FDR method. We noticed within the same cell type, pathway enrichments could sometimes be driven by a similar shared set of genes. We therefore calculated the Jaccard similarity coefficient of each pairwise enrichment signal (leading edge genes driving the high vs low responder difference), within each cell type, and adjusted intracellular correlation effect sizes such that they reflected “shared latent information” (SLI) by subtracting the Jaccard similarity index from the Spearman correlation coefficient ρ:

$$SLI = \rho - \frac{A \cap B}{A \cup B}$$

For example, given enriched pathways A and B within a cell type, if at one extreme, these two pathways are driven by the same exact shared 10 leading edge genes, the Spearman ρ of their normalized expression would be equal to 1, yet this apparent correlation is arbitrary since the two pathways reflect the same genes. However, the shared latent information would be equal to 0 because the Jaccard similarity of the two sets is also equal to 1 (since the leading edge genes from the enrichments are also the same). The remaining correlation strength reflected by SLI thus represents the phenotypic coupling of intracellular states across individuals due to another latent factor besides artifactually sharing similar driving genes. For inter-cellular correlations between two distinct cell types, we do not subtract the Jaccard similarity of gene content from ρ as we consider the same genes to be distinct signals when measured in different cell types. We further constructed a sub network from a subset of cell types forming the high responder baseline setpoint. Only correlations with adjusted p values < 0.05 were retained and a weighted undirected network was constructed using igraph, retaining only the strongest links above the median weight (with weights incorporating the SLI metric described above). Each node (high responder cell phenotype) was also correlated across individuals with the day 7 fold change of a gene expression signature[39] reflective of plasmablast activity derived from bulk microarray data from the same subjects and select high degree nodes were highlighted in the text.

Single-cell mixed-effect models of gene expression

Single cell mixed effects models were used to test the early kinetics of baseline states enriched above and select AS03 associated signatures within innate subsets.

Early kinetics of baseline associated cell phenotypes. Each cell type specific phenotype enriched in high vs low responders in the aggregated linear model described above were scored in single cells from subjects on day 0 and day 1 as the average expression of the specific leading edge genes enriched in high vs low responders. The single cell module scores were fitted with a linear mixed model for each cell type to 1) re-test the baseline association at the single cell level and 2) to test their post vaccination effect size within the same cell subset. These models estimated variation at the single-cell level instead of at the individual donor aggregated level. Otherwise they correspond to similar models

as described above fit using lme4, with a donor random intercept, without voom weights. Two models were tested with highly concordant effect sizes, 1) a parsimonious model of time relative to vaccination with a donor random effect and 2) a more complex model including the time relative to vaccination, the number of cells per individual sample for a given cell type, age, sex and a donor random effect. Normalized expression of each module was standardized within each protein based subset by subtracting the mean and dividing by the standard deviation of the module score across the cell type. After fitting models, the baseline high vs low responder effect and the day 1 vs baseline effect sizes and standard errors across subsets was calculated using the emmeans package with a custom contrast. All models were checked for convergence criteria.

AS03 specific innate cell subcluster expression of select modules. Monocyte and mDC combined protein + mRNA joint sub-clusters from the adjuvant cohort were tested for expression of AS03 specific modules derived from the aggregated model and gene set enrichment. The cell type specific leading edge genes were scored as above and fitted using the formula module score ~ 0 + timepoint + subcluster + subcluster:timepoint + (1|subjectid). This model was fitted across all monocytes to allow the vaccination effect to vary by subcluster while modeling variation across donors. The vaccination effect on module expression conditional on subcluster and the day 1 vs baseline effect was calculated using the emmeans package.

AS03 Innate cell sub cluster association with vaccination

To test the association of innate cell sub cluster with vaccination effects, we fitted an aggregated binomial mixed effects model. The model formula *n/total ~ time*cluster + (1|subjectid)* was fit using lme4 with the glmer function (family = 'binomial') and *weights* parameter equal to the total number of cells from each donor. This model enabled estimation of the proportion p of cells in each cluster c from each subject S belonging to each timepoint t accounting for within-donor replicated cells (i.e., pseudoreplication) in each cluster taking the form:

$$logit(p_c) = \beta_{0s} + c\beta_{1p} + t\beta_{2p} + ct\beta_{3p} + \varepsilon_p$$

$$\beta_{0s} = \gamma_0 + S_0$$

$$S_0 \sim N(0, {\tau_p}^2)$$

$$\varepsilon_p \sim N(0, \sigma_\varepsilon^2)$$

The log odds of a given sub cluster being increased in frequency post vaccination were calculated using the emmeans package.

Monocyte differentiation and perturbation pseudotime analysis

To construct a combined monocyte differentiation and perturbation single cell map we used the DDR tree algorithm with monocle2. The trajectory was constructed using the genes that changed as a function of time (q value <0.15 using the differentialGeneTest in monocle, ribosomal genes and genes expressed in less than 15 cells removed). The DDRtree algorithm was implemented using the monocle function reduceDimension with arguments *residualModelFormulaStr* = subjectID and *max_components* = 2 and pseudotime calculated with function orderCells. Independently of the genes used to construct the trajectory we then tested the genes from the mixed effects model of vaccination effects from monocytes (specific leading edge genes from 'reactome interferon signaling', 'GO IL6 PRODUCTION', 'reactome IL4 and IL13 signaling', 'HALLMARK inflammatory response', 'KEGG JAK STAT signaling') for branch dependent differential expression using the *BEAM* function from monocle. Select genes were highlighted and categorized based on their expression dynamics along real time and pseudotime.

CyTOF cell stimulation of high and low responder baseline cells

Samples were thawed in a 37°C water bath and washed twice with warmed complete media with Universal Nuclease (Pierce) added. Cells were then washed a final time and resuspended in complete media. 1 million cells per condition were added to individual wells and rested in a tissue culture incubator for 2 hours (37°C, 5% CO_2). Samples were then stimulated with either PMA/Ionomycin ((final concentration [10 ng/mL])/([1µg/mL]); Sigma-Aldrich), LPS (final concentration [1µg/mL]; Sigma-Aldrich), IFN-a (final concentration [10,000U/ml], PBL Assay Science), or left unstimulated. After 15 minutes at 37°C, samples were fixed with paraformaldehyde (2.2% PFA final concentration) for 10 minutes at 25°C. Samples were washed twice with Maxpar Barcode Perm Buffer (1X concentration; Standard Biotools). Samples were then barcoded with Cell-ID 20-Plex Pd Barcoding Kit (Standard Biotools) and

incubated at 25°C for 30 minutes. Samples were then washed twice with Maxpar Cell Staining Buffer (Standard Biotools) and combined into corresponding barcoded batches of 5 samples (4 conditions per sample) and washed a final time with Maxpar Cell Staining Buffer. Samples were then stained with a titrated antibody-panel for extracellular markers (Supplementary Table) for 30 minutes at 25°C. After staining, the cells were washed twice with Maxpar Cell Staining Buffer and permeabilized in methanol (Fisher Scientific) overnight at -80°C. The next day, samples were washed twice with Maxpar Cell Staining Buffer, and stained with a titrated panel of antibodies for intracellular signaling markers (Supplementary Table) at 25°C for 30 minutes. Samples were then washed twice with Maxpar Cell Staining Buffer, and labeled with Cell-ID Intercalator Ir ([1:2000] in Maxpar Fix-Perm Buffer; Standard Biotools) overnight at 4°C. The following day, samples were washed twice with Maxpar Cell Staining Buffer and resuspended in 500µL freezing media (90% FBS (Atlanta Biologicals) + 10% DMSO (Sigma-Aldrich), and stored at -80°C until acquisition. The day of acquisition, samples were thawed and washed twice with Maxpar Cell Staining Buffer and then once with Cell Acquisition Solution (Standard Biotools) before being resuspended in Cell Acquisition Solution supplemented with 10% EQ Four Element Calibration Beads at a concentration of 6 x 10^5 cells/mL (to approximate 300 events/sec). Samples were acquired on the Helios system (Standard Biotools) using a WB Injector (Standard Biotools). After acquisition, samples were normalized and debarcoded using the CyTOF Software's debarcoder and normalization tools (Standard Biotools). The panel and protocol were adapted for use at CHI from the Stanford HIMC. The phosphor markers driving the stimulated phenotype and responding cells were automatically defined using the HDStIM R package[316]. The median phosphomarker intensity for each individual sample and cell type and stimulation was calculated and modeled with a mixed effects model adjusting or batch and using a random effect for donor ID. The difference in fold change between unstimulated and stimulated cells was calculated using a custom contrast with the emmeans package.

Code availability

All code to replicate the analysis in this paper including all (Figures is available in the following repository: https://github.com/NIAID/fsc (not yet public). The computational framework used in this paper is available as an R package:

https://github.com/MattPM/scglmmr. This package was created to implement mixed effects models at the aggregated and single cell level from single cell genomics perturbation experiment data with repeated measures multi individual nested group designs.

4 CONTRASTING AUTOIMMUNE AND TREATMENT EFFECTS REVEALS BASELINE SET POINTS OF IMMUNE TOXICITY FOLLOWING CHECKPOINT INHIBITOR TREATMENT

This work is posted on BioRxiv

doi.org/10.1101/2022.06.05.494592

4.1 Abstract

Immune checkpoint inhibitors (ICIs) have changed the cancer treatment landscape, but severe immune-related adverse events (irAEs) can trigger life-threatening autoimmunity or treatment discontinuation. Uncovering immune phenotypes associated specifically with irAEs but not antitumor immunity could help mitigate treatment discontinuation and improve clinical outcomes. We carried out simultaneous transcriptome and surface protein profiling of blood immune cells from thymic cancer patients before and after treatment with the anti-PD-L1 antibody avelumab. All patients had antitumor responses, yet a subset developed severe myositis. Our analytical approach disentangled phenotypes linked to treatment responses versus irAEs and identified a temporally stable, pre-treatment immune set point associated with irAEs consisting of correlated innate and adaptive cell phenotypes, including genes downstream of mTOR in T-cell subsets. Together these findings suggest pre-treatment biomarkers of irAEs in thymic

cancer patients and raise the prospect of therapeutically dampening autoimmunity while sparing antitumor activity in cancer patients treated with ICIs.

4.2 Introduction

Immune checkpoint inhibitors (ICIs) have demonstrated durable benefit and improved survival in a subset of patients with advanced cancers[328]. However, this therapeutic benefit comes with a risk of immune-related adverse events (irAEs), common side effects of checkpoint inhibitor therapy ranging in frequency between around 50–90% depending on the type of cancer and checkpoint inhibitor[329–331]. These autoimmune reactions can be life-threatening, and can affect almost any organ system, with the most common symptoms being rash, pruritus, fatigue, and diarrhea[331]. While a majority of irAEs can be safely managed by discontinuing ICI treatments and/or giving low-dose steroids, some patients require high-dose steroids or anti-cytokine agents[332,333] which can decrease the antitumor effect of ICIs. Patients experiencing mild or moderate irAEs can be re-challenged with ICIs under close monitoring[334]; however, the risk of developing a subsequently fatal irAE often precludes continuation of treatment for patients developing severe autoimmunity. There is thus an urgent need for unbiased identification of molecular phenotypes associated with irAE risk to help inform potential biomarkers and treatment strategies to dampen autoimmune effects while sparing antitumor immunity[335].

Factors contributing to autoimmunity versus antitumor immunity in patients receiving ICIs remain unclear[336]. Immune inhibitory receptors targeted by these drugs play essential roles in maintaining self-tolerance, as documented in patients with germline mutations affecting these receptors and in transgenic mouse models lacking immune checkpoint inhibitory receptors[337–339]. Inhibition of negative feedback on immune activation by ICIs may thus cause autoimmune reactions in cancer patients by exacerbating pre-existing clinical or subclinical autoimmunity by increasing the probability of loss of immune tolerance[340]. IrAE rates are higher in patients treated with dual ICIs, yet single-agent ICI treatment is sufficient to cause autoimmunity in around 20% of patients[341,342]. Individual variations in baseline (i.e., pre-treatment) immune status (or "set points"[115,278,343]) may, for example, provide different levels of buffering (or pre-disposition) to develop adverse events. For example, a single "hit" to certain

regulatory pathways might be sufficient to cause pathology in some (e.g., those with less buffering capacity[344]) but not all patients. Identifying baseline pre-treatment molecular signatures and states associated with irAE outcomes could uncover biomarkers of immune toxicity with which to select patients for treatment and inform potential treatment interventions. A recent study shed light on the local reaction of T cells at the onset of irAE-related colitis, finding cycling T cells and alterations in T regulatory cells associated with irAEs[345]. Another report suggested baseline activated CD4 memory T-cell abundance could serve as a biomarker of post-treatment severe irAEs[92]. Despite these advances, previous unbiased systems-level analyses often used profiling approaches that have limited cellular resolution. Furthermore, statistical assessment of differences between signatures associated with irAEs and antitumor responses is lacking, but is critical for understanding the delicate interplay and shared mechanisms between ICI-induced autoimmunity and antitumor immunity. Biomarkers that specifically mark irAEs but not antitumor immunity could help in the development of interventional strategies with minimal impact on the efficacy of ICI therapies.

4.3 Study Design

We set out to contrast treatment-associated and irAE-associated immune system states by profiling peripheral immune cells of patients with metastatic thymic cancer before (baseline) and after administration of the anti-PD-L1 antibody avelumab (at the time or irAE development, or its equivalent in patients not developing irAE). We chose to study irAEs in thymic cancer for the disease's stable tumor cell-intrinsic property (low tumor mutation burden), good response to ICIs, and high incidence of irAEs[346,347]. Prior studies in other cancers using cytometry[348] and single-cell RNA sequencing[349,350] investigated responses to ICIs, yet the transcriptional state and phenotype of well-resolved immune cell populations before and after treatment are understudied, particularly involving contrasting treatment- and irAE-associated effects. We addressed these gaps by using CITE-seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing), a multimodal technique combining surface protein phenotyping and transcriptome profiling simultaneously in single cells, followed by mixed-effects modeling to identify cell subset-specific signatures associated with the development of irAE but not with clinical outcome. Our CITE-seq antibody panel targeted 82 surface proteins and included 4 isotype controls, as previously described[115]. Nine patients were chosen for CITE-seq analysis; all had clinically similar antitumor activities based on

RECIST (Response Evaluation Criteria in Solid Tumors). While no patients had clinically observable autoimmune disease at baseline, five individuals developed myositis after an average of two doses of avelumab. Paired peripheral blood mononuclear cell (PBMC) samples from baseline (pre-treatment) and at the onset of irAEs post-avelumab (two cycles post-avelumab for the non-irAE group) were used for analysis. Our dataset included more than 190,000 cells from 18 PBMC samples, with two timepoints per patient (Fig 4.1a) and a median of 10,804 cells per sample (Fig 4.2a-c).

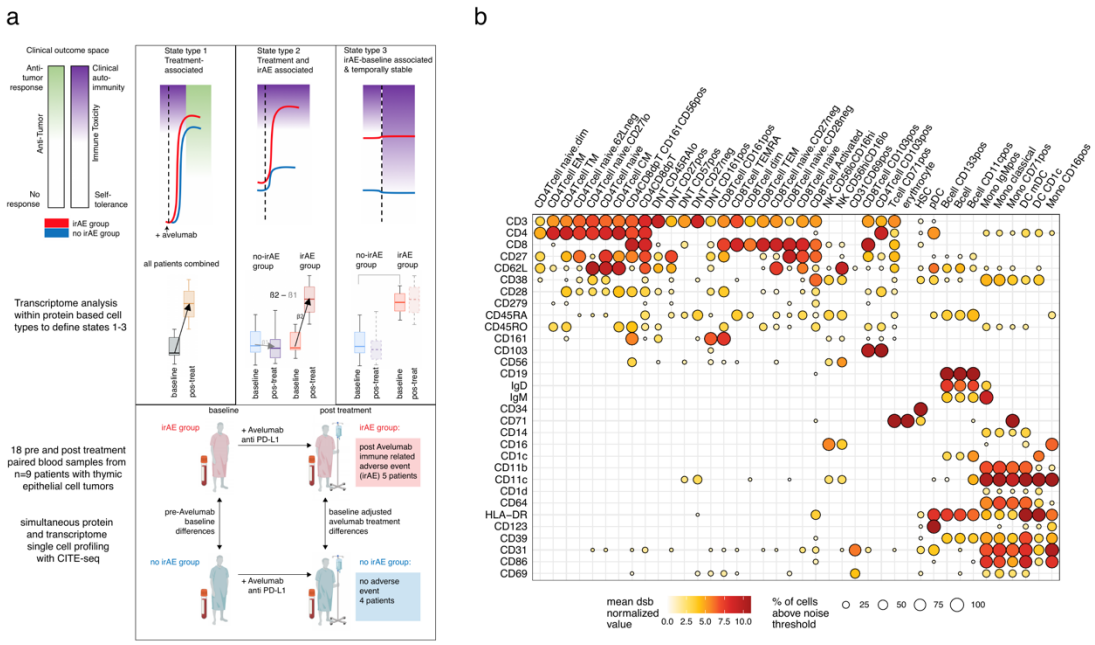


Figure 4.1 Multimodal single-cell analysis deconvolves transcriptome states associated with irAEs and ICI treatment within protein immune phenotypes

a. Top: hypothesized schematic illustrating different types of immune cell states and how those measured parameters reflect the clinical phenotypes: the antitumor response effects (white-blue) and autoimmune toxicity (green-red). Red lines represent the group of patients developing irAEs after treatment and blue represents those without irAEs. Cell states from left to right: State type 1: perturbed by treatment across all patients which can be associated with antitumor effects but also with autoimmune toxicity; State type 2: increased post-treatment in the irAE group (these could reflect a higher fold change in the irAE group or oppositely regulated states); State type 3: baseline differences in the irAE group exhibiting temporal stability over the course of treatment, which is associated with irAEs but not treatment effects. Middle: transcriptome comparisons carried out within protein clusters to identify different cell states corresponding to the states above. Bottom: study scheme devised to define cell state differences above: eighteen PBMC samples from nine patients with thymic cancers were profiled at baseline and post-avelumab treatment; five of them developed an irAE (myositis) post-treatment and the other four patients did not develop an irAE. PBMC

samples collected before treatment (baseline) and post-avelumab (at the onset of irAE and matched time points in the non-irAE group) were profiled using CITE-seq with a panel of 82 antibodies. b. CITE-seq surface protein expression map of PBMCs. Circle color is the mean dsb 'denoised' and normalized protein level; the scale of dsb values can be interpreted as the number of standard deviations above background noise. Circle size is the percentage of cells in the cluster that express the protein above the expression-positive cutoff of 3.5.

4.4 High dimensional protein-based immune cell phenotyping

Defining cell clusters and subsets with surface protein alone allowed us to identify cell type based on well-studied surface markers (Fig 4.1b, Fig 4.2d), thereby separating transcriptome measurements from cell type identity. This facilitated improved interpretation of transcriptome differences between outcome groups within cell clusters that were defined with statistically independent (protein) information from transcriptome data. We clustered cells using spectral clustering based on the denoised expression level of 82 surface proteins. This procedure identified 43 cell clusters spanning major cell lineages, including subsets of B cells, monocytes, dendritic cells (DCs), natural killer cells, and T cells (Fig. 4.1b, Fig. 4.2d). The substantial number of antibodies in our panel for marking T-lymphocyte phenotypes and cell states revealed significant heterogeneity within CD8+ and double negative (CD3+CD4–CD8–) T-cell subsets, with most of these clusters/phenotypes detected across donors (Fig. 4.2d).

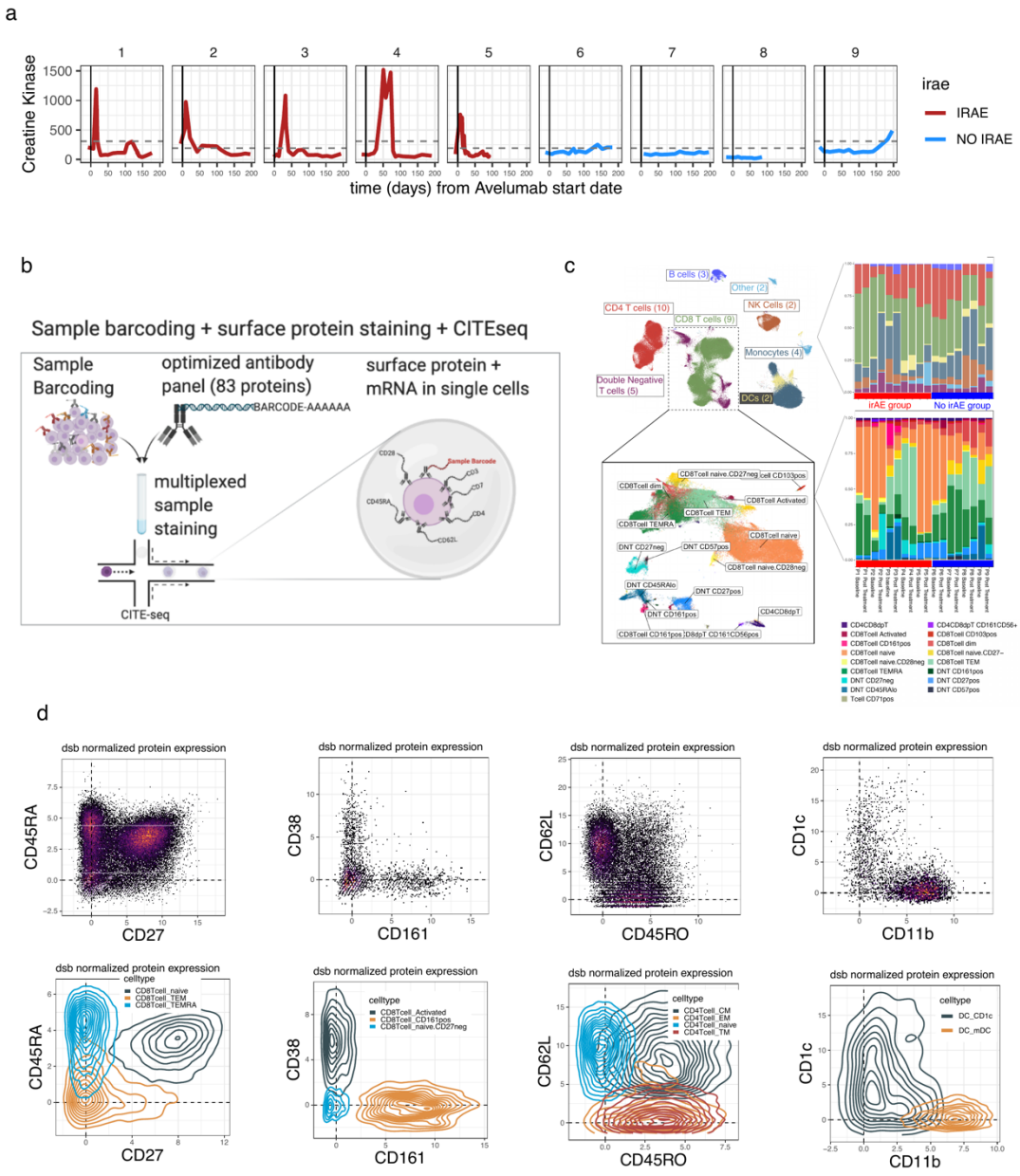


Figure 4.2 Clinical and protein based immune cell clustering details

a. Blood creatine kinase (CK) levels (y-axis) vs. time from the initiation of avelumab treatment in patients profiled with CITE-seq b. Sample multiplexing and CITE-seq experiment scheme. c. Left–Uniform manifold approximation projection (UMAP) of PBMC colored by main immune cell lineages with a subset of the map expanded (inset) containing the T cell subsets indicated in the box. Right–the distribution of the number of cells per subset for the clusters shown (top, main lineage, bottom, the T cell subsets shown in the bottom UMAP plot). d. CITE-seq dsb normalized and denoised surface protein expression from single cells in bi-axial plots with the corresponding cells in density plots colored by the spectral clustering annotation. The protein phenotypes align with known canonical cell types which enhances the interpretability of mRNA states associated with the clinical outcomes.

4.5 Statistical modeling of avelumab treatment and toxicity effects between patient groups

Our analysis approach focused on defining cell states associated with irAEs decoupled from treatment response within immune cell types defined by protein. To this end, we applied statistical contrasts to identify changes of cell functional states due to three major effects: 1) ICI treatment effects–pre- vs. post-avelumab treatment effects shared across all subjects; 2) ICI-associated irAE effects–the difference in pre- vs. post-treatment effect between irAE and non-irAE groups; and 3) baseline effects–differences in cell state prior to avelumab treatment between groups (analogously shown in Fig. 4.1a). By subtracting treatment effects (both effects 1 and 2) from baseline differences, we further focused on baseline cell states associated with impending irAEs exhibiting temporal stability over the course of treatment; these cell states were thus uncoupled from ICI response effects. To accommodate our experimental design containing patients with repeated measurements nested in groups, we used weighted mixed-effects models at the single-cell level and on pseudo-bulk data aggregated within each cluster to model variations across donors over time and between outcome groups (see Methods). We tested enrichment of a pre-specified list of gene modules based on our hypothesis of pathways that could tune immune states related to irAEs and response in addition to carrying out unbiased analysis of 50 MSigDB Hallmark pathways.

4.6 Defining immune checkpoint inhibitor treatment effects

We first assessed cell type-specific avelumab treatment effects (Fig. 4.1a- State type 1) by identifying, within each cell cluster above, the transcriptional differences between post- vs. pre-treatment across patients. Transcriptional signatures of T-cell activation, interferon pathways, PD-1 signaling and T-cell exhaustion were elevated within multiple T-cell subsets (Fig. 4.3a). Activated CD38++ CD8+ T cells and naïve CD8+ T cells had the highest number of enriched pathways. In the CD38++ subset, upregulation of T-cell activation and cell cycle signals included genes CDC25B, CDC27, MCM3, and CSK2 (Fig. 4.3a). Upregulation of GZMA, OAS1, IFITM2, LCK, MKI67, and PDCD1 in this subset driven by elevated activation protein CD38 was consistent with a proliferating effector phenotype (Fig. 4.3b). Interestingly, cell cycle states were enriched in the opposite direction (downregulated post-avelumab) in several other T-cell subsets including CD8+ TEMRA and CD8+ naïve T cells (Fig. 4.3a). Genes

driving pathway enrichments tended to be mutually exclusive among different cell clusters/subsets (as defined by the Jaccard similarity of the "leading edge" genes for each gene set/pathway; see Methods) (Fig. 4.4a). The presence of cell cycle signatures in CD38++ effector T cells post-avelumab treatment is reminiscent of phenotypes revealed in prior studies, e.g., those focusing on the dynamics of T-cell changes during ICI treatment[350,351] and a study of ICI-induced colitis[345]. Together, our and others' observations suggest that a proliferation signature of peripheral CD8+ effector T cells is coupled to ICI treatment responses.

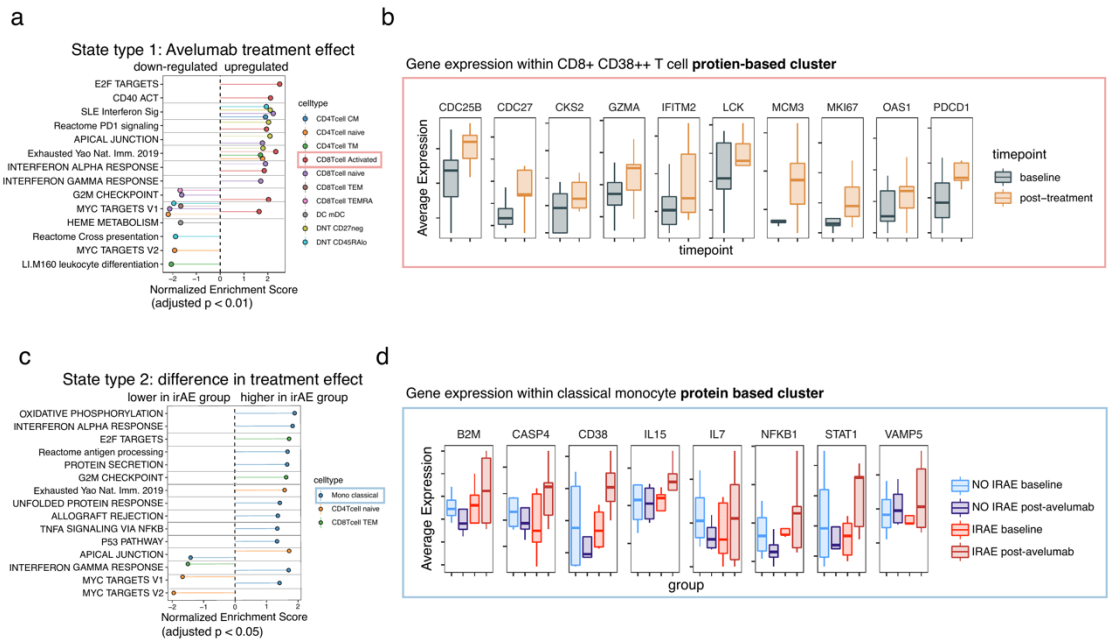


Figure 4.3 Avelumab treatment effects across individuals and specific to irAEs

a. State type 1: treatment effects : gene set enrichment based on genes ranked by the pre- vs. post-avelumab treatment effect from donor-weighted pseudobulk models fit within protein based subsets. b. Selected genes from the CD38++CD8+ effector T-cell cluster in leading-edge genes of the enriched pathways shown. c. State type 2: difference in treatment effects between the irAE and non-irAE group: gene set enrichment. d. Selected genes from the classical monocyte cluster with a treatment effect upregulated in the irAE group – genes include those with oppositely regulated directions (CD38 mRNA) or genes only perturbed in the irAE group (IL15).

4.7 Defining treatment effects unique to immune related adverse events

While ICI may trigger qualitatively similar responses in different patients, the quantitative extent of these responses may differ between irAE and non-irAE patients. To evaluate this possibility, we identified treatment response-associated irAE effects by

comparing the difference in post-treatment vs. baseline fold changes between the irAE and non-irAE groups (Fig. 4.1a, State type 2); this analysis identified 35 enrichments involving 14 cell types. Within classical monocytes in particular, interferon transcriptional signatures were highly enriched (Fig. 4.3c) and leading-edge genes in these pathway enrichments tended to be mutually exclusive (Fig. 4.4b). IL15 and interferon-simulated genes (ISGs) B2M, CD38 and STAT1 were oppositely regulated between groups from baseline to post-treatment (upregulated in the irAE group and downregulated in the non irAE group). These further suggested that interferon-responsive monocytes were activated in the irAE group after treatment with avelumab (Fig. 4.3d). This IFN response in monocytes that is associated with irAEs may phenocopy the elevated interferon response states seen in inflammatory diseases such as lupus[352], and has also been observed in patients with myositis[353]. In addition, cell cycle/proliferation signals were enriched post-treatment in the irAE compared to the non-irAE group in CD8 effector memory T cells (Fig. 4.3c).

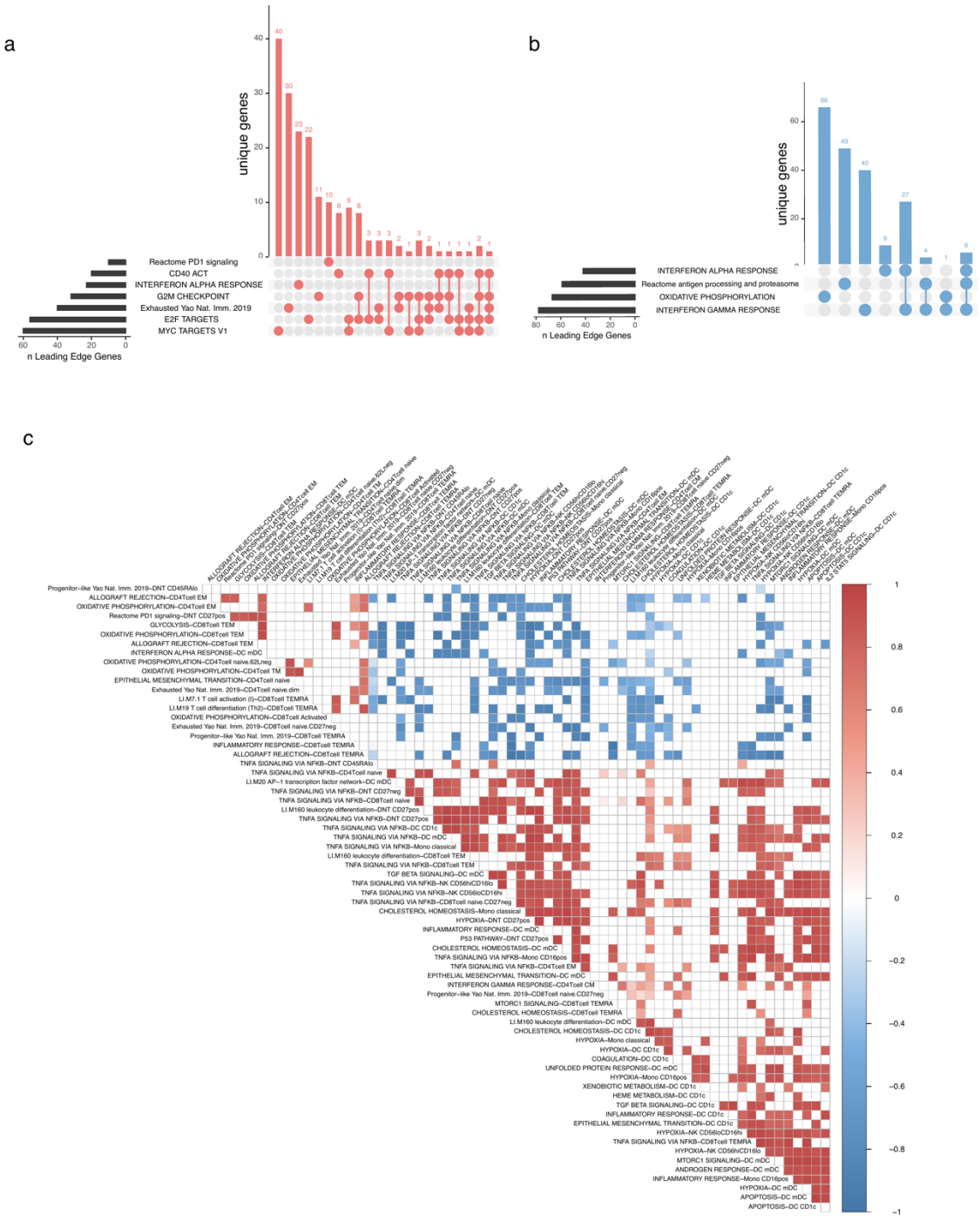


Figure 4.4 Shared information in molecular phenotypes related to avelumab treatment and irAEs

a. UpSet plot of the intersection of leading edge genes for cell state type I (treatment effect) in CD38++ CD8+ T cells shown in Fig 4.3a. b. As in (a) for classical monocyte enrichments shown in Fig 4.3c. c. An expanded version of the baseline cell state correlation map shown in fig 4.5d. Sample level baseline pseudobulk expression correlations of temporally stable baseline states associated with development of later irAE. Each box represents the Pearson correlation coefficient (two sided) of donor pseudobulk data with all correlations with FDR adjusted p value < 0.02 not shown.

4.8 A baseline metabolic transcriptional signature is associated with post-treatment irAEs independent of treatment effects

We next searched for signatures of baseline (prior to treatment) immune states in patients who developed irAEs post-treatment. We focused on signatures uncoupled from avelumab treatment effects. We first identified baseline states associated with post-treatment irAEs, then subtracted enrichment signals associated with avelumab treatment (see Methods). This procedure resulted in a map of temporally stable cell type-specific signatures, or "set points", associated with the post-treatment development of irAEs independent of treatment effects (i.e., those defined by Fig. 4.3a-d, State type 1 and 2 in Fig. 4.1a). Inflammatory and metabolic signatures including mTOR and TNFα pathway genes were enriched within multiple cell subsets (Fig. 4.5a). DCs have known roles in modulating autoimmune and antitumor responses[354] and both CD1chigh and CD1clow DCs in the irAE group appeared to have an elevated inflammatory signature at baseline, e.g., TGFB, TNFA, and inflammatory response pathway enrichments. They also displayed potential enhanced tissue migratory capacity given the enrichment of epithelial-to-mesenchymal transition genes, including CD44, VIM, VCAN, THBS1, and SDC4 (Fig. 4.5a). These primed DC subsets were also elevated for several metabolic-related transcript differences, for example, involving mTORC1 signaling, hypoxia, and cholesterol homeostasis. Some of these inflammatory and metabolic signatures are also shared by CD8+ T cells, in particular memory cells re-expressing CD45RA (CD8+ TEMRA), which displayed elevated TNF signaling, hypoxia, cholesterol homeostasis, and mTOR signatures in the irAE group (Fig. 4.5a). By design, these phenotypes we uncovered were not enriched in the irAE group after avelumab treatment, where innate immunity and inflammatory signatures such as those associated with IFNs were more specific to CD14+ classical monocytes (Figs. 4.3c,d).

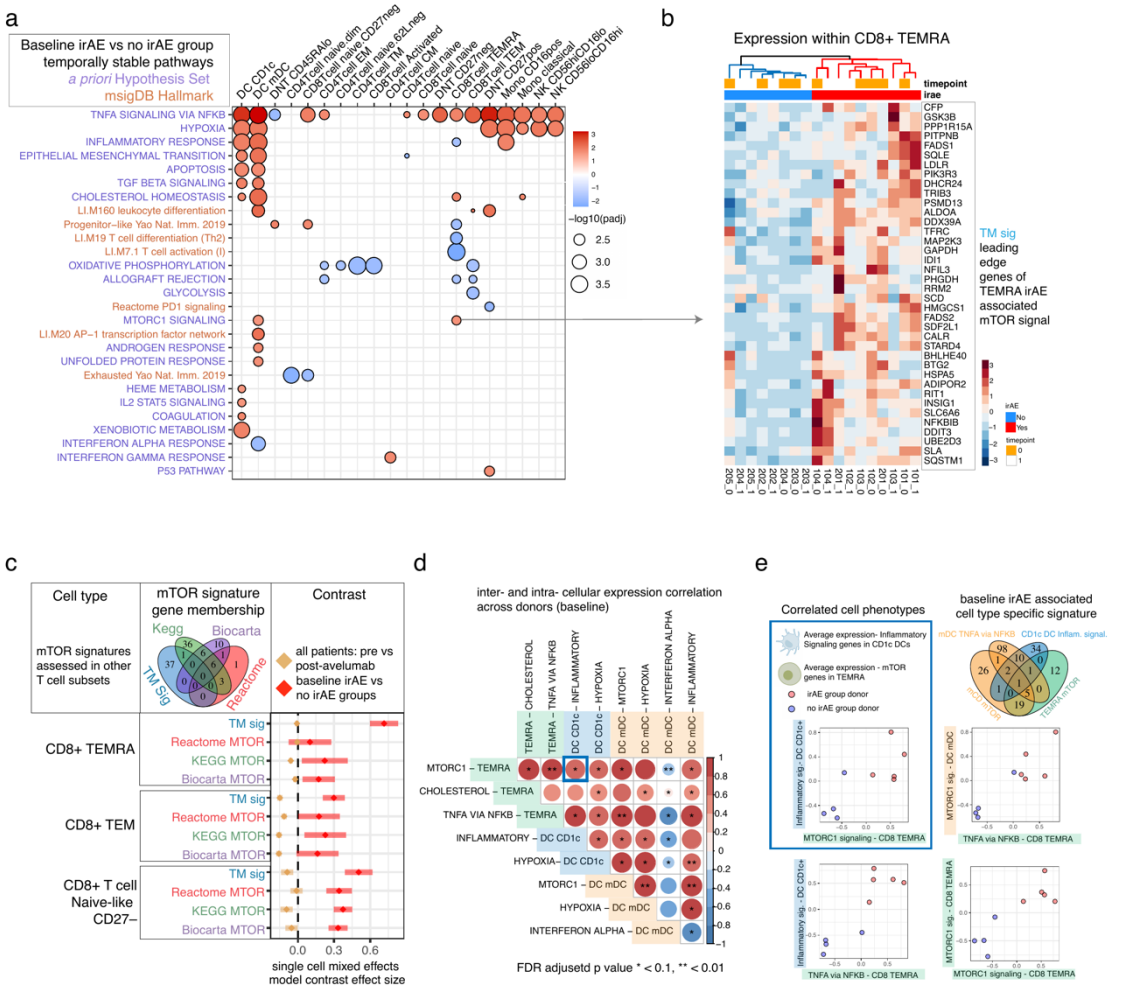


Figure 4.5 Baseline immune set points associated with irAEs

a. Baseline gene set enrichment map based on genes ranked by weighted pseudobulk model effect size within each cluster comparing patients with eventual irAEs to those without irAEs. Red and blue indicate pathways positively and negatively enriched respectively in patients before avelumab treatment who developed an irAE after receiving avelumab. The pathway names (y-axis) highlighted in blue are from the MSigDB hallmark collection, and pathways in dark orange are curated gene sets based on a pre-study defined hypothesis (Supplementary Table 3). b. Average single cell expression of leading-edge genes from the Hallmark mTOR pathway baseline enrichment associated with the irAE group within the TEMRA cell cluster. Samples from both time points are ordered according to hierarchical clustering with complete linkage. c. The coefficient corresponding to baseline irAE vs. no-irAE (red) and the fold change across donors (tan) from a single-cell mixed-effects model of other mTOR signatures across CD8 T-cell subsets. Tan lines with a coefficient effect size near 0 are temporally stable; pathways with an effect size above 0 are associated with irAEs; error bars are 95% confidence intervals of the contrast applied to mixed-model fits. d. Sample-level baseline pseudobulk expression correlations of temporally stable baseline cell states associated with development of irAEs after avelumab treatment. Each box represents the Pearson correlation coefficient (two-sided) of donor pseudobulk data with the FDR adjusted p-value (FDR adjustment across all temporally stable baseline enrichments, subset of states shown in correlation matrix) shown with asterisks. e.

Scatterplots of average donor expression for selected inter- and intracellular correlations shown in d.

4.9 Correlated cell-state phenotypes underlie baseline set point signatures of irAEs

Given the critical role of the mTOR pathway in tumorigenesis[355] and autoimmunity[356], we further examined the genes driving mTOR pathway enrichment within CD8 TEMRA cells (Fig. 4.5a). These leading-edge mTOR genes (mTOR-LE) naturally clustered in all samples, independent of time points, into two clusters segregated by irAE status (Fig. 4.5b). The independence from time points confirmed that our procedure identified temporally stable enrichment signals that were stable between the pre- and post-treatment time points. Expression of mTOR-LE genes was elevated in thymic cancer patients compared to both the non-irAE group and healthy donors (n=20) assessed with the same CITE-seq panel within gated CD8 TEMRA cells (Fig. 4.6a-c). mTOR-LE genes such as SLC2A1, GAPDH, FADS1, FADS2, LDLR, and ADIPOR2 (Fig. 4.5b) suggested this enrichment may reflect a metabolic state downstream of mTOR, since these genes are involved in glucose and lipid metabolism. Therefore, we further tested 6 distinct mTOR signatures covering different aspects of the pathway from public databases (Fig. 4.6c-e) by repeated differential expression models using a k-cell permutation approach (see Methods). All mTOR signals as well as the TNF pathway were consistently enriched in the irAE group in CD8 TEMRA, (Fig. 4.6d), suggesting the mTOR-LE signal may have reflected an immune state controlled in part by upstream mTOR signaling. We next wondered if this state could be shared by other cell types. Using a more sensitive model accounting for variation at the single cell level and modeling expression of the 6 mTOR pathways defined above (see Methods) revealed mTOR signatures elevated at baseline in the irAE group within double-negative (DNT: CD4–CD8–), double-positive, and CD8+ subsets including CD8+ TEMRA, TEM, CD27– naïve-like cells (Fig. 4.5c), while CD4 subsets were not enriched for mTOR pathways. Consistent with temporal stability of the original CD8 TEMRA mTOR-LE signature, baseline elevation in the irAE group was uncoupled from treatment effects in these subsets (Fig. 4.5c, tan estimate and errors near 0). Notably, mTOR signatures were not upregulated in the irAE group within the activated CD8+ CD38++ T-cell cluster which had a post-treatment phenotype overlapping with

antitumor responses[357] (see above Fig. 4.3a,b). To investigate molecular identity of these protein-based cell types independent of the irAE group differences, we integrated healthy donors' and thymic cancer patients' CITE-seq data into a joint CD8 T-cell map (Fig. 4.7a-d). CD8 TEMRA were distributed across multiple clusters including cluster 1, which was enriched for mTOR-LE genes and expressed genes controlling terminal effector fate HOPX[358],34 GZMH,[359] and the transcription factor ZEB2[360] (Fig. 4.7c).

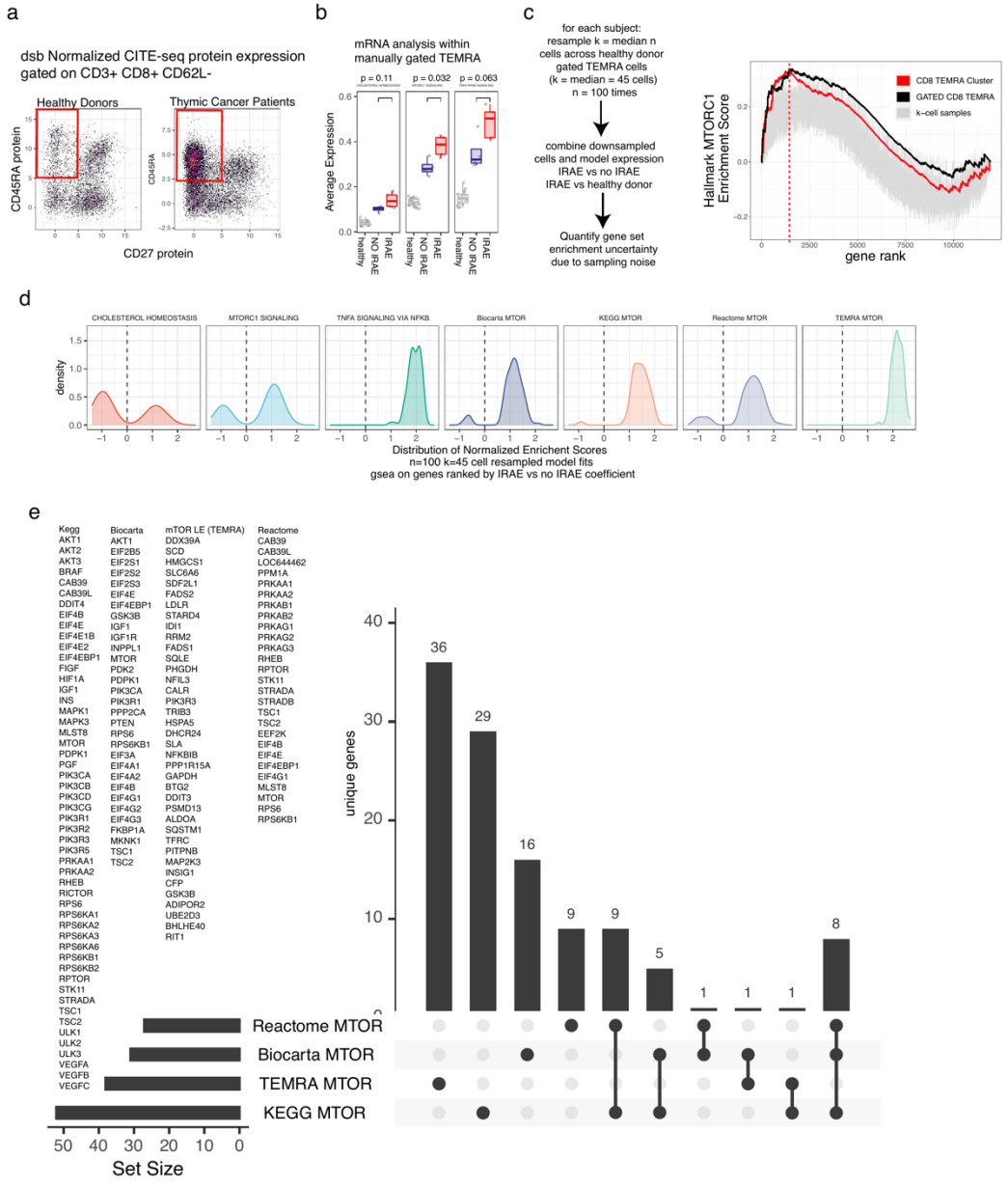


Figure 4.6 Robustness assessment of baseline signatures of irAEs

a. Manually gating CD8 TEMRA cells from healthy donors and thymic cancer patients as CD3+ CD8+ CD62L- CD45RA+ CD27- based on dsb normalized protein expression. b. Average expression of the leading mTOR-LE gene signature in the manually gated subsets. c. Robustness assessment of mTOR signature enrichment in

irAE vs non-irAE thymic cancer patients and healthy donors; k=45 random cells were re-sampled from each donor to form a downsampled pseudobulk library the resampling procedure was repeated 100 times each followed by a full reanalysis of the weighted pseudobulk model with gene set enrichment based on effect size for the contrast indicated. Grey lines reflect the n=100 down sampled analysis with enrichment from the analysis using the full data (all cells from each donor) shown as the black line and the original enrichment signal from the TEMRA cluster in red. d. As in c; showing the full distribution of normalized enrichment scores across the k-cell permutation testing procedure. e. UpSet plot of mTOR signatures tested in (d) with gene membership in each module shown.

To further characterize shared information between distinct baseline irAE-associated enrichment signals, we correlated baseline expression of enriched pathway leading-edge genes across subjects both within and between cell types (Fig. 4.5d, Fig. 4.4c). Supporting our hypothesis that CD8 TEMRA mTOR captured a more global irAE-associated metabolic state, as reflected by mTOR's elevation across different effector and naive CD8 subsets in the irAE group (Fig. 4.5c), CD8 TEMRA mTOR-LE expression correlated with other irAE-associated metabolic and inflammatory signaling pathways across subjects (top row of Fig. 4.5d). For example, the CD1c DC inflammatory signaling and TEMRA mTOR phenotypes were correlated across donors (inset in blue in Fig. 4.5e). The level of the TEMRA mTOR signal was correlated with TEMRA TNF signaling (Fig. 2e) and this TNF signal was associated with the level of innate-cell mTOR and inflammatory signaling (Fig. 4.5d,e). The irAE baseline innate inflammatory and metabolic phenotype appeared distinct from states related to interferon tone, as mDC interferon signaling was negatively enriched in the irAE group and negatively correlated with metabolic and inflammatory states across subsets. Together, these correlated cell phenotypes suggest stable inter- and intracellular rewiring of inflammatory and metabolic states comprising a shared immune set point of patients primed toward development of autoimmunity after treatment with avelumab.

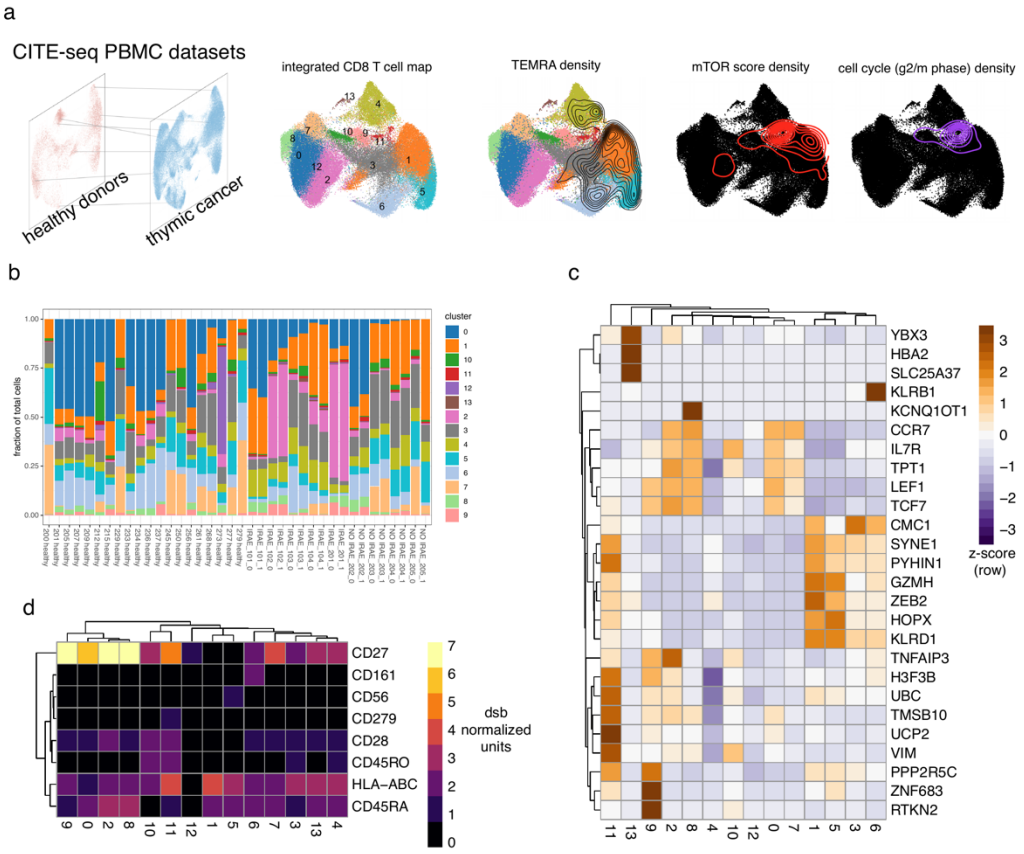


Figure 4.7 Integrated T cell embedding with healthy donor and thymic cancer T cells

a. CD8 T cells from the n=20 healthy donor cohort were used as a reference dataset on which to project thymic cancer patient cells to form an integrated CITE-seq healthy and cancer T cell map. Cells are colored by integrated assay clustering using Seurat based on 2000 genes regularized Pearson residuals values. The density of the TEMRA cells from the thymic cancer cohort are shown overlaying the integrated mRNA based clusters. The density of cells with mTOR score > 3 absolute deviations from the median mTOR signature score is shown second from right. The right-most UMAP is as above, but for the density of g2m phase score. b. The proportion of cells from each donor belonging to each cluster in the integrated clustering. c. Differential expression of markers between clusters based on regularized Pearson residuals with donor effect regressed out, ROC test implemented in Seurat. d. dsb normalized protein expression in clusters as in (c).

4.10 Assessing irAE-associated T-cell signature in tissue-localized T cells associated with ICI-induced colitis

The circulating CD8+ T-cell set point signature we identified may phenotypically overlap with those found in tissues associated with adverse immune reactions. We investigated this further by assessing our CD8+ T-cell signatures in single-cell RNA

sequencing data obtained from a published study of CD8+ T cells isolated from colonoscopy biopsies of healthy donors (n=8) and patients with melanoma with (n=8) and without (n=6) active ICI-induced colitis[345]. We focused on eight CD8+ T-cell clusters defined by unsupervised clustering (Fig. 4.8a,b); three of these CD8 clusters were specific to the CD8+ T cells isolated from colitis lesions (Fig. 4.8c) as identified by a mixed effects frequency model comparing the colitis to the no colitis group, all of which had an effector phenotype based on mRNA expression (Fig. 4.8d). We further examined cluster 4 ("effector 1") and cluster 10 ("effector 2") as these had sufficient numbers of cells after aggregation across subjects (see Methods). Within these T-cell clusters, donors with colitis had higher relative expression of the mTOR-LE gene signature compared to T cells from either healthy donors or ICI-treated melanoma patients without colitis (Fig. 2.8e). Interestingly, BHLHE40, a member of our mTOR-LE signature, is the gene with the most significant differential expression between ICI-induced colitis vs. non-colitis in cluster 10 (effector cluster 2, data not shown). This gene is a crucial regulator of cytokine production associated with autoimmune responses[361], which is consistent with the notion that the mTOR-LE gene signature reflects a "poised" metabolic/inflammatory phenotype.

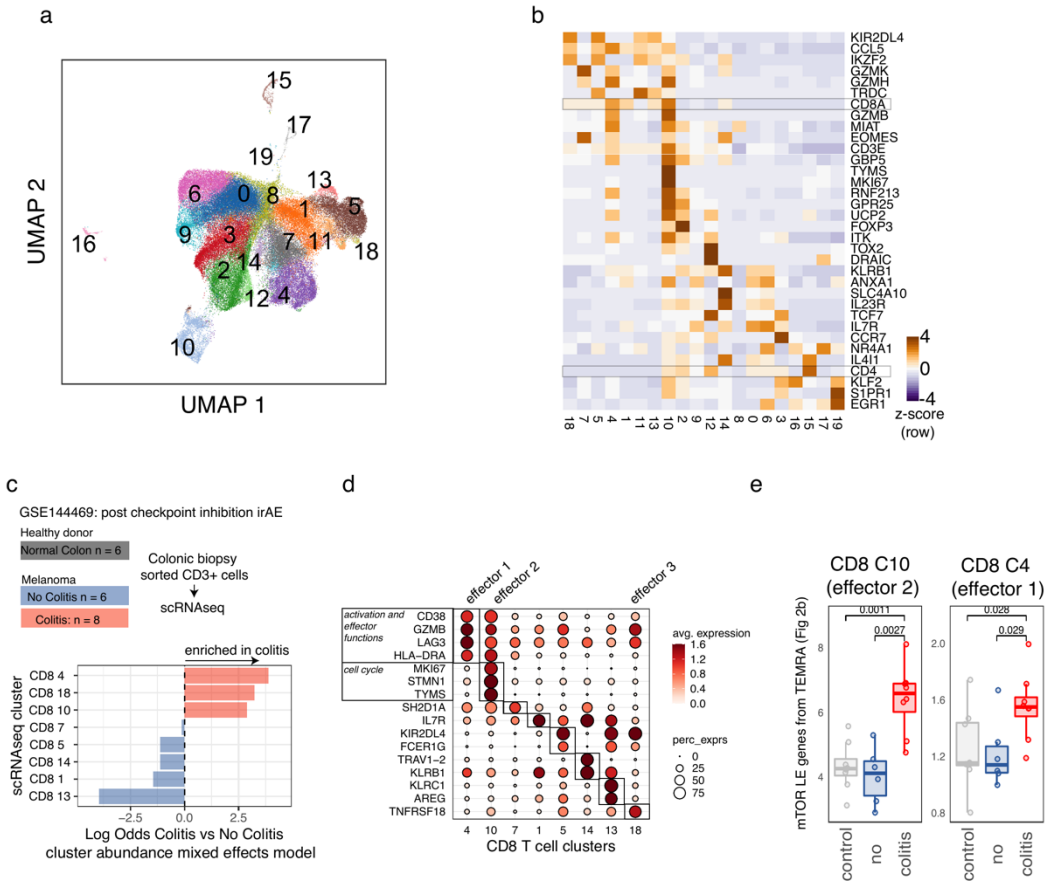


Figure 4.8 Evaluation of blood irAE signatures in checkpoint inhibitor induced colitis colonic tissue T cells

a. UMAP projection of CD3+ cells from melanoma and healthy colonic biopsies; colors indicate clusters derived from graph-based clustering based on 30 principal components from regularized Pearson residuals of 2000 genes. b. Average cluster expression of select genes used for annotation in Luoma et al. c. Structure of experiment data from Luoma et. al. 2020 (GSE144469) which profiled patients treated with combined checkpoint inhibitors who went on to have suspected colitis that was either confirmed on biopsy with overt colitis (red) vs. no evidence of colitis (blue) and healthy colon biopsies (grey) with CD3+ cells FACS sorted followed by single-cell RNA-seq. Bottom shows association testing using an aggregated binomial generalized linear mixed model of the association of cells from each cluster with the colitis vs. no colitis groups. d. Expression of selected differentially expressed genes for each cluster of colonic T cells from a one cluster vs. all ROC test (Seurat). e. The average colitis T-cell expression of the mTOR-LE gene signature within effector CD8 T-cell clusters 10 and 4 across donors.

4.11 Discussion of Chapter 4 results

In this work we identify a set of highly interpretable multimodal molecular states associated with ICI response and adverse events. We found multiple cell functional

states linked to ICI response were also likely coupled to those involved in irAEs. However, patients with post-treatment irAEs shared a common baseline immune set point, reflecting elevated inflammatory tone and metabolic differences across the innate and adaptive immune systems. Our analysis suggested that this baseline set point may be tuned by common upstream regulators such as mTOR, which is known to regulate metabolic states such as hypoxia[362]. Given the role of mTOR inhibitors as both antitumor agents and suppressants of autoimmunity, our results provide the rationale for evaluating the concurrent use of mTOR inhibitors with ICI in an attempt to diminish the risk of developing irAEs while preserving the antitumor effects of ICI. Intriguingly, a case report of a renal allograft patient with melanoma treated with an ICI and an mTOR inhibitor found the antitumor effect could be preserved while the autoimmune toxicity could be limited[363]. However, further studies are needed to confirm these observations.

Our study has several limitations. Multimodal single-cell profiling of more than 190,000 cells created a high-resolution map of cell states, but the number of patients included in our study is limited and, due to experimental constraint at the time of this experiment, samples were split across two batches. We have previously found staining batch has limited impact on technical effects in CITE-seq data[115,250]. However, assessment of batch effects across sample groups was limited due to the small sample size confounding in this dataset. Future single-cell analysis of irAEs could include additional subjects to assess the generalizability of the molecular states we derived herein. The generalizability of these findings to other types of cancers and ICIs could also be assessed in future work, although we did find overlap in signals within tissue autoimmunity from melanoma patients treated with different ICIs. Finally, it will be interesting to link differences in immune cell states detected from blood with those at the tissue level from sites of involvement by the irAEs. While these clones would be difficult to trace in humans, lineage tracing mouse model systems could be informative in studying the origins of these cells. Together, our dataset, analysis, and curated results can serve as both a framework and a rich source of hypothesis-generating data to inform future precision immunotherapy research on biomarkers and treatment strategies for irAEs.

4.12 Methods – Chapter 4

Clinical/sample collection

Patients with advanced thymic cancers were enrolled in clinical trial NCT03076554 approved by the NCI's Institutional Review Board and received avelumab, an anti-PD-L1 antibody, every two weeks. PBMC samples were collected before starting therapy, at the end of every treatment cycle, and at the onset of irAEs. No patients had any history of autoimmune disease prior to treatment.

Multiplexed CITE-seq single-cell transcriptome and protein profiling

Cells were thawed in RPMI with 10% FBS and washed and stained in 1xPBS with 0.04% BSA. CITE-seq was performed as previously described in Kotliarov et al. using the same antibody panel. Donor cells were stained with sample barcoding antibodies[98], washed, and pooled into a single tube; two staining batches were used to accommodate a greater number of samples than available barcode antibodies. Although a single batch design was planned using lipid indexing[324], a subset of samples had red cells noticeable in the PBMC prep; we therefore used HTO staining in 2 batches due to unknown effect of residual RBC membranes on LMO staining (Supplementary Table 4). Pooled cells were stained with a concentrated optimized panel of 86 antibodies (including 4 isotype controls; anti-mouse (rather than anti-human) CD206 was incorrectly included in the panel and not considered in the analysis). The stained cell pool was then washed and prepared according to the 10X Genomics cells partitioned across eight lanes of the 10X Genomics chromium microfluidic instrument per staining batch. Sequencing libraries were prepared using the 10X Genomics 3' assay with version 3 reagents. Antibody-derived tag (ADT) libraries from sample barcode antibodies and surface phenotyping antibodies were prepared according to the publicly available protocol on cite-seq.com. Sequencing was performed on an Illumina NovaSeq system.

Normalizing and denoising CITE-seq protein levels and protein-based clustering

After sample demultiplexing and doublet removal based on sample barcoding antibodies, CITE-seq surface protein data were normalized and denoised using dsb[100] to correct protein-specific background noise using ADT reads in empty droplets and correct technical cell-to-cell variations using isotype controls/models fitted to each cell.

Default dsb algorithm parameters were used (denoise.counts = TRUE, use.isotype.controls = TRUE). The normalized values were then batch corrected using limma. Single cells were clustered using a Euclidean distance matrix formed from the normalized protein values as input to spectral clustering using Seurat version 3.1.5[272]. A total of 44 cell clusters were annotated based on protein expression.

Analysis of aggregated transcriptome data within protein-based clusters

Gene expression counts were aggregated into a pseudo-bulk library within each protein-based cluster by adding counts for each sample x cell type into a summed count matrix, and cell types without representation (e.g., donor-specific clusters) were excluded from analysis. The aggregated counts for the n=18 samples across each cell type were normalized using the trimmed means of M values method[364] and genes were retained which had a pooled count per million above 3 across sufficient samples based the edgeR filterByExprs function. Filtering genes in a cell type-specific manner removed genes from analysis unexpressed by a given cell type (e.g., genes specific to a different lineage) and ensured assumptions of the model to derive precision weights[73] used to account for variations in sample quality/library size were met, i.e., the log count per million vs. fitted residual square root standard deviation had the expected monotonically decreasing trend within each cell type (see below).

Estimating subject and group-level effects within protein-based clusters

Target estimates of statistical analysis were treatment and group-level transcriptional effects within the protein-based clusters defined above. Models were fitted to single-cell and aggregated data (see below). To assess these effects, we contrasted fitted values of fixed and mixed effects linear models of gene expression within protein clusters. A contrast matrix L was constructed with a single combined factor variable group.time corresponding to irAE outcome group and time point relative to treatment with levels 1 = irAE baseline, 2 = irAE post-avelumab, 3 = no irAE baseline, 4 = no irAE post-avelumab (columns, below). The matrix was used to make the following comparisons (rows) based on fitted model values (see below) 1) ICI treatment effects–across all subjects, 2) ICI-associated irAE effects–the fold change difference between groups and 3) baseline effects–the baseline difference between the irAE and non-irAE group.

$$L = \begin{bmatrix} -0.5 & 0.5 & -0.5 & 0.5 \\ -1 & 1 & 1 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix}$$

The first two rows of contrast matrix above were applied to estimate 1) the coefficients for the treatment effect across all donors and 2) the difference in fold changes between the irAE and non-irAE groups from mixed-effect model fits. Mixed effects models on aggregated data to estimate treatment effects across donors and fold change differences between groups. Estimation of avelumab treatment effect across all donors and the difference in treatment effects between irAE groups was modeled with a mixed-effects model including a varying effect for subject ID to model variation in baseline expression. Models were fit using the formula f1 = gene ~ 0 + group.time + (1|subjectID) and fit models using the "dream" method[71] as in Chapter 3.

Modeling baseline states associated with development of irAE

The third row of the contrast matrix above was used to estimate baseline differences using a fixed-effects model with limma with the function lmFit using voom[73] precision weights as above in a fixed-effects model. The Empirical Bayes moderated t statistics for each gene comparing the irAE group to the non-irAE group were calculated using the limma[70] eBayes function. After gene set enrichment (see below "Enrichment testing of hypothesis set and unbiased pathways in model contrasts"), we defined the subset of these baseline states associated with later irAEs which exhibited temporal stability over the course of treatment. The pathways enriched in the irAE group with adjusted p values < 0.01 were further filtered by removing any enrichments evidence of kinetic change (including weak evidence). For each enriched pathway within each cluster, if either the treatment across donors or the irAE-associated treatment effect enrichments (see below "Estimating avelumab treatment effects across donors and between groups") had adjusted p values of 0.1 or less and were either positively or negatively enriched, these were considered kinetically altered by treatment for the purpose of filtering the baseline signals. These kinetically altered signals were subtracted from the baseline enrichments in a cell type-specific fashion, and the remaining enriched baseline pathways associated with development of post-treatment irAEs were considered temporally stable states.

Single-cell mixed-effect models

The same formula f1 above (see "Mixed effects models on aggregated data to estimate treatment effects across donors and fold change differences between groups") was used in a linear mixed model on expression of gene modules within single cells in specific T-cell subsets. The model estimated variation at the single-cell level instead of at the individual donor aggregated level and otherwise corresponds to the same model formula as described above without voom observational weights in the error term. Gene expression of each gene *g* in each cell *i* was normalized log transformed with library size scaling factors using the Seurat function NormalizeData() with *normalization.method* = 'LogNormalize' to implement the transformation:

$$\log\left(1 + \frac{10^5 \times UMI_{i,g}}{\sum_i UMI}\right)$$

Average expression of gene modules/pathways was then calculated for each module for each single cell and standardized within each protein-based subset by subtracting the mean and dividing by the standard deviation of the average score. Models were fit using the R package lme4[72] and the treatment effect across donors, and the baseline difference between irAE and no irAE groups was estimated using the contrast matrix *L* above with the emmeans package[365]. Models were checked for convergence criteria and no models were flagged as having singular fits.

Enrichment testing of hypothesis set and unbiased pathways in model contrasts

To test enrichment of pathways based on the estimated gene coefficients corresponding to the three effects defined above, we performed gene set enrichment analysis using the fgsea package[274] using 250,000 permutations of the ranked gene list to form null distributions for p values; genes were ranked based on the empirical Bayes moderated t-statistic for the baseline comparison of irAE status or with the raw t-statistic for mixed-effect models comparing treatment effects over time. Two gene sets were assessed: first a hypothesis set of modules curated from the Li et al. Blood Transcriptional modules[88]

MSigDB Hallmark[366] Reactome[367] and pathways curated from literature[368–370] were tested for enrichment (Supplementary Table 3); the full MSigDB Hallmark pathways were tested independently. The Jaccard similarity of enrichments within cell types was calculated using the geneOverlap[371] package.

k-cell permutation profiling of CD8 T-cell signatures and enrichment

To assess the robustness of gene set enrichments, T_{EMRA} cells were manually gated based on dsb normalized CITE-seq protein expression of CD3,CD8,CD45RA, and CD27. The same cells were gated from the previously published data on 20 healthy donors from Kotliarov *et al.,* and the average expression of the mTOR-LE genes was compared across thymic cancer irAE groups and healthy donors using a non-parametric Wilcoxon rank test. To account for variability in the number of cells per donor in both manually gated and unsupervised T_{EMRA} clusters, we quantified stability of enrichment to cell sampling variations. We re-ran the pseudobulk baseline differential expression model (as described above "Modeling baseline states associated with development of irAE") 100 times with libraries constructed from random k-cell samples (without replacement) of 45 cells from each donor. The k value of 45 was chosen as it was the median number of cells in the group (healthy donors) with the lowest number of gated T_{EMRA} across all donors. Pseudobulk libraries were constructed and differential expression testing of the irAE vs. non-irAE groups and healthy donors was carried out as above using limma. Genes within each k sample were tested for enrichment using two complementary methods with highly concordant results. 1) genes were filtered for testing based on the pseudobulk expression profile of the k cell pool with a minimum of 3 counts per million based on the design matrix as above. 2) The same genes as in the original T_{EMRA} cluster were fit with limma regardless of their expression status in the k cell pool. Genes were then ranked by empirical Bayes t statistic comparing irAE vs. non-irAE or irAE vs. healthy donors (not shown and highly concordant with irAE vs. non-irAE, as expected based on the average expression profiles in Supplemental Fig. 3b). Gene set enrichment was assessed using 250,000 permutations of the gene rank list for the null distribution, as described above.

Integrated analysis of healthy donors and thymic cancer patients

Seurat version 3.1.5 was used to integrate healthy donor and thymic cancer PBMC data using healthy donors as the reference dataset. Regularized Pearson residuals[233] were used to normalize data for integration with a covariate for subject. Integrated data were clustered using 30 principal components. Differential expression was compared between integrated clusters with an ROC test in Seurat (FindAllMarkers, test.use = 'roc'). Immune cells from cancer patients often clustered more distinctly; however, this shared state map based only on mRNA helped further define mRNA substates irrespective of the group comparisons described above.

Analysis of colonic T cells from patients with and without colitis following checkpoint inhibitor treatment from Luoma *et al.* 2020

We reanalyzed the colonic T-cell data from Luoma *et al.* (GSE144469) which included three patient groups: healthy donors, patients treated with ICIs with subsequent colitis (irAE group) and without colitis (non-irAE group). T cells were isolated from the site of the colitis lesion in the irAE group. T-cell single-cell mRNA data were clustered using 30 principal components from 2000 variable genes (FindVariableFeatures with selection.method = 'vst') using Seurat version 3. Differential expression of genes between clusters was carried out using a one-cluster vs. all-ROC classifier implemented in Seurat. Cluster association with irAE group status was carried out with an aggregated binomial mixed-effects model to estimate the proportion p of cells in each cluster c from each subject S belonging to each group g accounting for within-donor replicated cells (i.e., pseudoreplication) in each cluster using the same mixed effects binomial count model in Chapter 3.

Code availability

Analysis code and documentation to reproduce this work is available in the

repository: https:// https://github.com/niaid/irAE_manuscript

5 CONCLUDING REMARKS

More detailed discussions related to specific results of each chapter are given in the discussion sections above. At the beginning of undertaking this thesis work, the initial report of the CITE-seq method and two reports of sample multiplexing methods had just been published. We scaled these technologies to large scale systems immunology studies. Experimentally, we utilized both genetic and antibody based barcoding / demultiplexing to extend single cell profiling complex experiment designs with many individuals with different perturbation timepoints (see methods of chapter 3). We then deconvolved noise sources in the protein data modality and used this information to develop a scalable normalization method, dsb, which has been adopted by the community[46,372]. Analytically we further developed a framework for integrating human population, cell subset, and single cell variations multimodal single cell data. We applied both of these methods to clinical cohorts described in chapter 2 and 3 which helped identify multicellular perturbation states and immune setpoints linked to treatment response. We have also successfully applied these approaches to several other projects since then[250]. A major goal of top down systems immunology we help advance through these approaches is the identification of precise baseline and perturbation states within immune cell subsets and their correlated activities associated with desired responses.

With these approaches, we provided the first evidence for very detailed molecular circuits forming a "naturally adjuvanted" baseline immune system phenotype. Many questions remain unknown, for example, what determines the setpoint of high

responders? Could differential activation of endogenous retroviruses activate innate immune cells to help program the setpoint? Given the apparent role of tonic interferon signaling in the baseline setpoints related to vaccination response, it could be fruitful to investigate differences in endogenous retroviruses within immune subsets found in discarded sequencing reads. Since the models we used to define baseline setpoints adjusted for age sex and batch, these factors likely are not major determinants of the molecular phenotypes we identified. It is still possible that other intrinsic variations like genetics played a role in defining the baseline states. Out other recent work not discussed in this thesis has identified persistent changes that can be induced by vaccination on the scale of multiple months. Is it possible that high responders had recovered from a viral or bacterial illness mirroring this process in the months prior to the vaccination study, leading to a more poised baseline state?

Conceptualizing these baseline immune states as a "setpoint" borrows the term from system control theory which presupposes a homeostatically regulated desired goal state. Whether we should use such descriptions to define the global properties of immune system should be further debated. For example, does the notion of a setpoint represent an ontological property of immune system behavior, or is it a more tautological representation? Levin and colleagues have written on the usefulness of defining such "goal state" endpoints in top down systems biology modeling[51]. If we consider the time scales and environmental diversity over which the immune system must function, singular setpoints leading to deterministic responses would not give enough flexibility for the immune system to adjust to local environments. The setpoints we identified are comprised of correlated modular activity "poised" within multiple innate cell subsets. Perhaps the "naturally adjuvanted" baseline state leading to better vaccine responses were dictated by adaptations to local microenvironments dictated by pathogen diversity and antigen encounter frequency. These may globally tune degrees of "sentinel" capacity of innate cells lasting on the order of months. In the context of our work defining *shared* setpoints between autoimmunity and vaccination responses, if the determinants of the setpoint could be identified, they could be potentially controlled and tuned to improve autoimmune disease outcomes.

With respect to autoimmunity, ongoing work applies many of the subset level "layer 3" statistical models to sorted subsets from vasculitis, inflammatory bowel disease, lupus, and idiopathic pulmonary fibrosis. These datasets build on prior work linking cell

subset level immune states during active disease to clinical outcomes[52,373] but include many more defined immune subsets. These datasets represent an ideal opportunity to test the pathways unperturbed by immunomodulating treatment, since they are collected during active disease prior to treatment. Whether the molecular activities of these disease associated immune subsets overlaps with states induced by vaccination in healthy individuals, adverse events in patients following checkpoint inhibition, or have similarities with the multicellular baseline circuitries identified here will be interesting to investigate.

1. NOSSAL, G. J. V. & LEDERBERG, J. Antibody Production by Single Cells. *Nature* **181**, 1419–1420 (1958).

2. Brack, C., Hirama, M., Lenhard-Schuller, R. & Tonegawa, S. A complete immunoglobulin gene is created by somatic recombination. *Cell* **15**, 1–14 (1978).

3. Davis, M. M. & Bjorkman, P. J. The T cell receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).

4. Janeway, J. How the immune system works to protect the host from infection: A personal view. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7461–7468 (2001).

5. Iwasaki, A. & Medzhitov, R. Control of adaptive immunity by the innate immune system. *Nat. Immunol.* **16**, 343–353 (2015).

6. Kasturi, S. P. *et al.* Programming the magnitude and persistence of antibody responses with innate immunity. *Nature* **470**, 543–550 (2011).

7. Crotty, S. A brief history of T cell help to B cells. *Nat. Rev. Immunol.* **15**, 185–189 (2015).

8. Mahnke, Y. D., Brodie, T. M., Sallusto, F., Roederer, M. & Lugli, E. The who's who of T-cell differentiation: Human memory T-cell subsets. *Eur. J. Immunol.* **43**, 2797–2809 (2013).

9. Roederer, M. *et al.* The genetic architecture of the human immune system: A bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).

10. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

11. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

12. Zilbauer, M. *et al.* Genome-wide methylation analyses of primary human leukocyte subsets identifies functionally important cell-type-specific hypomethylated regions. *Blood* **122**, 52–60 (2013).

13. Uhlen, M. *et al.* A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science (80-.).* **366**, (2019).

14. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627-1640.e7 (2019).

15. Peters, J. E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS Genet.* **12**, (2016).

16. Watkins, N. A. *et al.* A HaemAtlas: Characterizing gene expression in differentiated human blood cells. *Blood* **113**, 1–9 (2009).

17. Westerhoff, H. V. & Palsson, B. O. The evolution of molecular biology into systems biology. *Nat. Biotechnol.* **22**, 1249–1252 (2004).

18. Davis, M. M. & Brodin, P. Rebooting Human Immunology. *Annu. Rev. Immunol.* **36**, 843–864 (2018).

19. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science (80-.).* **370**, (2020).

20. Tsang, J. S. Utilizing population variation, vaccination, and systems biology to study human immunology. *Trends Immunol.* **36**, 479–493 (2015).

21. Postow, M. A., Callahan, M. K. & Wolchok, J. D. Immune checkpoint blockade in cancer therapy. *J. Clin. Oncol.* **33**, 1974–1982 (2015).

22. Conroy, M. & Naidoo, J. Immune-related adverse events and the balancing act of immunotherapy. *Nat. Commun.* **13**, 1–4 (2022).

23. McDade, T. W. The ecologies of human immune function. *Annu. Rev. Anthropol.* **34**, 495–521 (2005).

24. Hill, D. L. *et al.* Immune system development varies according to age, location, and anemia in African children. *Sci. Transl. Med.* **12**, (2020).

25. Brodin, P. *et al.* Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. *Cell* **160**, 37–47 (2015).

26. Brodin, P. Immune-microbe interactions early in life: A determinant of health and disease long term. *Science (80-.).* **376**, 945–950 (2022).

27. Dowling, D. J. & Levy, O. Ontogeny of early life immunity. *Trends Immunol.* **35**, 299–310 (2014).

28. Germain, R. N. The art of the probable: System control in the adaptive immune system. *Science (80-.).* **293**, 240–245 (2001).

29. Pulendran, B. Systems vaccinology: Probing humanity's diverse immune systems with vaccines. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 12300–12306 (2014).

30. Hagan, T., Nakaya, H. I., Subramaniam, S. & Pulendran, B. Systems vaccinology: Enabling rational vaccine design with systems biological approaches. *Vaccine* **33**, 5294–5301 (2015).

31. De Jager, P. L. *et al.* ImmVar project: Insights and design considerations for future studies of 'healthy' immune variation. *Semin. Immunol.* **27**, 51–57 (2015).

32. Germain, R. N., Meier-Schellersheim, M., Nita-Lazar, A. & Fraser, I. D. C. Systems biology in immunology: A computational modeling perspective. *Annu. Rev. Immunol.* **29**, 527–585 (2011).

33. Germain, R. N. Will systems biology deliver its promise and contribute to the development of new or improved vaccines?: What really constitutes the study of "systems biology" and how might such an approach facilitate vaccine design. *Cold Spring Harb. Perspect. Biol.* **10**, (2018).

34. Brodin, P. & Davis, M. M. Human immune system variation. *Nat. Rev. Immunol.* **17**, 21–29 (2017).

35. Scheiermann, C., Kunisaki, Y. & Frenette, P. S. Circadian control of the immune system. *Nat. Rev. Immunol.* **13**, 190–198 (2013).

36. Dimitrov, S. *et al.* Cortisol and epinephrine control opposing circadian rhythms in T cell subsets. *Blood* **113**, 5134–5143 (2009).

37. Dopico, X. C. *et al.* Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nat. Commun.* **6**, (2015).

38. Lakshmikanth, T. *et al.* Human Immune System Variation during 1 Year. *Cell Rep.* **32**, (2020).

39. Tsang, J. S. *et al.* Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell* **157**, 499–513 (2014).

40. Carr, E. J. *et al.* The cellular composition of the human immune system is shaped by age and cohabitation. *Nat. Immunol.* **17**, 461–468 (2016).

41. Bergamaschi, L. *et al.* Longitudinal analysis reveals that delayed bystander CD8+ T cell activation and early immune pathology distinguish severe COVID-19 from mild disease. *Immunity* **54**, 1257-1275.e8 (2021).

42. Lyons, P. A. *et al.* Microarray analysis of human leucocyte subsets: The advantages of positive selection and rapid purification. *BMC Genomics* **8**, 1–12 (2007).

43. Gomes, T., Teichmann, S. A. & Talavera-López, C. Immunology Driven by Large-Scale Single-Cell Sequencing. *Trends Immunol.* **40**, 1011–1021 (2019).

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

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

46. Mimitou, E. P. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).

47. Swanson, E. *et al.* Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using tea-seq. *Elife* **10**, 1–38 (2021).

48. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).

49. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

50. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7426–7431 (2005).

51. Pezzulo, G. & Levin, M. Top-down models in biology: Explanation and control of complex living systems above the molecular level. *J. R. Soc. Interface* **13**, (2016).

52. McKinney, E. F., Lee, J. C., Jayne, D. R. W., Lyons, P. A. & Smith, K. G. C. T-cell exhaustion, co-stimulation and clinical outcome in autoimmunity and infection. *Nature* **523**, 612–616 (2015).

53. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science (80-.).* **345**, 1254665–1254665 (2014).

54. Brodin, P. The biology of the cell – insights from mass cytometry. *FEBS J.* **286**, 1514–1522 (2019).

55. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* **165**, 780–791 (2016).

56. Pope, S. D. & Medzhitov, R. Emerging Principles of Gene Expression Programs and Their Regulation. *Mol. Cell* **71**, 389–397 (2018).

57. Fowler, T., Sen, R. & Roy, A. L. Regulation of primary response genes. *Mol. Cell* **44**, 348–360 (2011).

58. Glass, C. K. & Natoli, G. Molecular control of activation and priming in macrophages. *Nat. Immunol.* **17**, 26–33 (2016).

59. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).

60. Zakrzewska, A. *et al.* Macrophage-specific gene functions in Spi1-directed innate immunity. *Blood* **116**, 1–11 (2010).

61. Bentebibel, S. E. *et al.* Induction of ICOS+CXCR3+CXCR5+ T H cells correlates with antibody responses to influenza vaccination. *Sci. Transl. Med.* **5**, 176ra32 LP-176ra32 (2013).

62. Rathmell, J. C., Townsend, S. E., Xu, J. C., Flavell, R. A. & Goodnow, C. C. Expansion or elimination of B cells in vivo: Dual roles for CD40- and Fas (CD95)-ligands modulated by the B cell antigen receptor. *Cell* **87**, 319–329 (1996).

63. Lewontin, R. C. The analysis of variance and the analysis of causes. *Int. J. Epidemiol.* **35**, 520–525 (2006).

64. Yuan, A. E. & Shou, W. Data-driven causal analysis of observational biological time series. *Elife* **11**, 1–49 (2022).

65. Park, Y. P. & Kellis, M. CoCoA-diff: counterfactual inference for single-cell gene expression analysis. *Genome Biol.* **22**, 1–23 (2021).

66. Piasecka, B. *et al.* Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E488–E497 (2018).

67. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

68. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

69. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 1–13 (2015).

70. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

71. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192–201 (2021).

72. Bates, D. *et al.* Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

73. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

74. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 273–282 (2011).

75. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, 47–52 (1999).

76. Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067–2075 (2005).

77. Candia, J. & Tsang, J. S. ENetXplorer: An R package for the quantitative exploration of elastic net families for generalized linear models. *BMC Bioinformatics* **20**, 1–9 (2019).

78. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 768 (2005).

79. Vallania, F. *et al.* Leveraging heterogeneity across multiple datasets increases cell-mixture deconvolution accuracy and reduces biological and technical biases. *Nat. Commun.* **9**, (2018).

80. Warsinske, H. C. *et al.* Assessment of Validity of a Blood-Based 3-Gene Signature Score for Progression and Diagnosis of Tuberculosis, Disease Severity, and Treatment Response. *JAMA Netw. open* **1**, e183779 (2018).

81. Lee, T. I. *et al.* Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science (80-.).* **298**, 799–804 (2002).

82. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.* **31**, 64–68 (2002).

83. Nachman, I., Regev, A. & Friedman, N. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* **20**, 248–256 (2004).

84. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).

85. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* **5**, 1–10 (2010).

86. Chaussabel, D. *et al.* A Modular Analysis Framework for Blood Genomics Studies: Application to Systemic Lupus Erythematosus. *Immunity* **29**, 150–164 (2008).

87. Altman, M. C. *et al.* Development of a fixed module repertoire for the analysis and interpretation of blood transcriptome data. *Nat. Commun.* **12**, 1–19 (2021).

88. Li, S. *et al.* Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).

89. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

90. Saelens, W., Cannoodt, R. & Saeys, Y. A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* **9**, (2018).

91. Olin, A. *et al.* Stereotypic Immune System Development in Newborn Children. *Cell* **174**, 1277-1292.e14 (2018).

92. Lozano, A. X. *et al.* T cell characteristics associated with toxicity to immune checkpoint blockade in patients with melanoma. *Nat Med* **28**, 353–362 (2022).

93. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).

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

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

96. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 1–12 (2017).

97. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

98. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 1–12 (2018).

99. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

100. Mulè, M. P. *et al.* Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nat. Commun.* **13**, 2099 (2022).

101. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).

102. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).

103. Lönnberg, T. *et al.* Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Sci. Immunol.* **2**, 1–12 (2017).

104. Gelman, A. & Stern, H. The difference between 'significant' and 'not significant' is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).

105. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).

106. Barreiro, L. B. *et al.* Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. *PLoS Genet.* **5**, (2009).

107. Von Bernuth, H. *et al.* Pyogenic Bacterial Infections in Humans with MyD88 Deficiency. *Science (80-.).* **321**, 691–696 (2008).

108. Zhang, S. Y. *et al.* TLR3 deficiency in patients with herpes simplex encephalitis. *Science (80-.).* **317**, 1522–1527 (2007).

109. Prugnolle, F. *et al.* Pathogen-driven selection and worldwide HLA class I diversity. *Curr. Biol.* **15**, 1022–1027 (2005).

110. Fumagalli, M. *et al.* Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *J. Exp. Med.* **206**, 1395–1408 (2009).

111. Thaventhiran, J. E. D. *et al.* Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature* **583**, 90–95 (2020).

112. Lettre, G. & Rioux, J. D. Autoimmune diseases: Insights from genome-wide association studies. *Hum. Mol. Genet.* **17**, 116–121 (2008).

113. Lee, J. C. *et al.* Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat. Genet.* **49**, 262–268 (2017).

114. Lee, J. C. *et al.* Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell* **155**, 57–69 (2013).

115. 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. *Nat. Med.* **26**, 618–629 (2020).

116. Patin, E. *et al.* Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors resource. *Nat. Immunol.* **19**, 302–314 (2018).

117. Lu, Y. *et al.* Systematic Analysis of Cell-to-Cell Expression Variation of T Lymphocytes in a Human Cohort Identifies Aging and Genetic Associations. *Immunity* **45**, 1162–1175 (2016).

118. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

119. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

120. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).

121. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

122. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

123. Li, Y. *et al.* A Functional Genomics Approach to Understand Variation in Cytokine Production in Humans. *Cell* **167**, 1099-1110.e14 (2016).

124. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).

125. Li, Y. *et al.* Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. *Nat. Med.* **22**, 952–960 (2016).

126. Jadidi-Niaragh, F. & Mirshafiey, A. Th17 Cell, the new player of neuroinflammatory process in multiple sclerosis. *Scand. J. Immunol.* **74**, 1–13 (2011).

127. Tesmer, L. A., Lundy, S. K., Sarkar, S. & Fox, D. A. Th17 cells in human disease. *Immunol. Rev.* **223**, 87–113 (2008).

128. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science (80-.).* **343**, (2014).

129. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science (80-.).* **343**, 1246980–1246980 (2014).

130. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–565 (2016).

131. Maier, L. M. *et al.* IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet.* **5**, (2009).

132. Lowe, C. E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.* **39**, 1074–1082 (2007).

133. Wong, H. S. *et al.* A local regulatory T cell feedback circuit maintains immune homeostasis by pruning self-activated T cells. *Cell* **184**, 3981-3997.e22 (2021).

134. Franco, L. M. *et al.* Integrative genomic analysis of the human immune response to influenza vaccination. *Elife* **2013**, 1–18 (2013).

135. Schadt, E. E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).

136. Klein, S. L., Jedlicka, A. & Pekosz, A. The Xs and Y of immune responses to viral vaccines. *Lancet Infect. Dis.* **10**, 338–349 (2010).

137. Liva, S. M. & Voskuhl, R. R. Testosterone Acts Directly on CD4 + T Lymphocytes to Increase IL-10 Production . *J. Immunol.* **167**, 2060–2067 (2001).

138. Khan, D. & Ansar Ahmed, S. The immune system is a natural target for estrogen action: Opposing effects of estrogen in two prototypical autoimmune diseases. *Front. Immunol.* **6**, 1–8 (2016).

139. Meier, A. *et al.* Sex differences in the Toll-like receptor-mediated response of plasmacytoid dendritic cells to HIV-1. *Nat. Med.* **15**, 955–959 (2009).

140. Cook, I. F. Sexual dimorphism of humoral immunity with human vaccines. *Vaccine* **26**, 3551–3555 (2008).

141. ter Horst, R. *et al.* Host and Environmental Factors Influencing Individual Human Cytokine Responses. *Cell* **167**, 1111-1124.e13 (2016).

142. Voigt, E. A. *et al.* Sex differences in older adults' immune responses to seasonal influenza vaccination. *Front. Immunol.* **10**, (2019).

143. Gaucher, D. *et al.* Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses. *J. Exp. Med.* **205**, 3119–3131 (2008).

144. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.* **10**, 116–125 (2009).

145. Furman, D. *et al.* Systems analysis of sex differences reveals an immunosuppressive role for testosterone in the response to influenza vaccination. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 869–874 (2014).

146. Park, J. E. *et al.* A cell atlas of human thymic development defines T cell repertoire formation. *Science (80-.).* **367**, (2020).

147. Yu, W. *et al.* Clonal Deletion Prunes but Does Not Eliminate Self-Specific αβ CD8+ T Lymphocytes. *Immunity* **42**, 929–941 (2015).

148. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: How selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2010).

149. Pawelec, G. Does the human immune system ever really become 'senescent'? *F1000Research* **6**, 1–7 (2017).

150. Fagnoni, F. F. *et al.* Shortage of circulating naive CD8+ T cells provides new insights on immunodeficiency in aging. *Blood* **95**, 2860–2868 (2000).

151. Goronzy, J. J. & Weyand, C. M. T-cell senescence and contraction of T-cell repertoire diversity - Catalysts of autoimmunity and chronic inflammation. *Arthritis Res. Ther.* **5**, 225–234 (2003).

152. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13139–13144 (2014).

153. Goronzy, J. J. & Weyand, C. M. Understanding immunosenescence to improve responses to vaccines. *Nat. Immunol.* **14**, 428–436 (2013).

154. Peters, M. J. *et al.* The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, (2015).

155. Ferrucci, L. & Fabbri, E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat. Rev. Cardiol.* **15**, 505–522 (2018).

156. Bektas, A. *et al.* Age-associated alterations in inducible gene transcription in human CD4+ T lymphocytes. *Aging (Albany. NY).* **5**, 18–36 (2013).

157. Kaczorowski, K. J. *et al.* Continuous immunotypes describe human immune variation and predict diverse responses. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6097–E6106 (2017).

158. Shen-Orr, S. S. *et al.* Defective Signaling in the JAK-STAT Pathway Tracks with Chronic Inflammation and Cardiovascular Risk in Aging Humans. *Cell Syst.* **3**, 374-384.e4 (2016).

159. Sobolev, O. *et al.* Adjuvanted influenza-H1N1 vaccination reveals lymphoid signatures of age-dependent early responses and of clinical adverse events. *Nat. Immunol.* **17**, 740–740 (2016).

160. Furman, D. *et al.* Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol. Syst. Biol.* **9**, 1–14 (2013).

161. Avey, S. *et al.* Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses. *Sci. Immunol.* **2**, eaal4656 (2017).

162. Avey, S. *et al.* Seasonal Variability and Shared Molecular Signatures of Inactivated Influenza Vaccination in Young and Older Adults. *J. Immunol.* **204**, 1661–1673 (2020).

163. Nakaya, H. I. *et al.* Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures. *Immunity* **43**, 1186–1198 (2015).

164. Thakar, J. *et al.* Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. *Aging (Albany. NY).* **7**, 38–52 (2015).

165. Hill, D. L. *et al.* Impaired HA-specific T follicular helper cell and antibody responses to influenza vaccination are linked to inflammation in humans. *Elife* **10**, 1–30 (2021).

166. Belkaid, Y. & Hand, T. W. Role of the microbiota in immunity and inflammation. *Cell* **157**, 121–141 (2014).

167. Davar, D. *et al.* Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients. *Science (80-.).* **371**, 595–602 (2021).

168. Vescovini, R. *et al.* Massive Load of Functional Effector CD4 + and CD8 + T Cells against Cytomegalovirus in Very Old Subjects . *J. Immunol.* **179**, 4283–4291 (2007).

169. Sylwester, A. W. *et al.* Broadly targeted human cytomegalovirus-specific CD4+ and CD8+ T cells dominate the memory compartments of exposed subjects. *J. Exp. Med.* **202**, 673–685 (2005).

170. Michaelis, M., Doerr, H. W. & Cinatl, J. The story of human cytomegalovirus and cancer: Increasing evidence and open questions. *Neoplasia* **11**, 1–9 (2009).

171. Freedman, B. Cytomegalovirus seropositivity and C-reactive protein have independent and combined predictive value for mortality in patients with angiographically demonstrated coronary artery disease. *Circulation* **104**, 1917–1923 (2001).

172. Pera, A. *et al.* CMV latent infection improves CD8+ T response to SEB due to expansion of polyfunctional CD57+ cells in young individuals. *PLoS One* **9**, 1–7 (2014).

173. OLD, L. J., CLARKE, D. A. & BENACERRAF, B. Effect of Bacillus Calmette-Guérin Infection on Transplanted Tumours in the Mouse. *Nature* **184**, 291–292 (1959).

174. Kaufmann, E. *et al.* BCG Educates Hematopoietic Stem Cells to Generate Protective Innate Immunity against Tuberculosis. *Cell* **172**, 176-190.e19 (2018).

175. de Laval, B. *et al.* C/EBPβ-Dependent Epigenetic Memory Induces Trained Immunity in Hematopoietic Stem Cells. *Cell Stem Cell* **26**, 657-674.e8 (2020).

176. Cirovic, B. *et al.* BCG Vaccination in Humans Elicits Trained Immunity via the Hematopoietic Progenitor Compartment. *Cell Host Microbe* **28**, 322-334.e5 (2020).

177. Myrianthefs, P. *et al.* Seasonal variation in whole blood cytokine production after LPS stimulation in normal individuals. *Cytokine* **24**, 286–292 (2003).

178. Temba, G. S. *et al.* Urban living in healthy Tanzanians is associated with an inflammatory status driven by dietary and metabolic changes. *Nat. Immunol.* **22**, 287–300 (2021).

179. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1125-1136.e8 (2016).

180. Hagan, T. *et al.* Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans. *Cell* **178**, 1313-1328.e13 (2019).

181. Furman, D. *et al.* Cytomegalovirus infection enhances the immune response to influenza. *Sci. Transl. Med.* **7**, 281ra43-281ra43 (2015).

182. Fuller, C. L. *et al.* Transcriptome analysis of human immune responses following live vaccine strain (LVS) Francisella tularensis vaccination. *Mol. Immunol.* **44**, 3173–3184 (2007).

183. Osman, M. *et al.* A third generation vaccine for human visceral leishmaniasis and post kala azar dermal leishmaniasis: First-in-human trial of ChAd63-KH. *PLoS Negl. Trop. Dis.* **11**, 1–24 (2017).

184. Gonçalves, E. *et al.* Innate gene signature distinguishes humoral versus cytotoxic responses to influenza vaccination. *J. Clin. Invest.* **129**, 1960–1971 (2019).

185. Ovsyannikova, I. G. *et al.* Gene signatures associated with adaptive humoral immunity following seasonal influenza A/H1N1 vaccination. *Genes Immun.* **17**, 371–379 (2016).

186. Anderson, J. *et al.* Molecular signatures of a TLR4 agonist-adjuvanted HIV-1 vaccine candidate in humans. *Front. Immunol.* **9**, 1–11 (2018).

187. Bucasas, K. L. *et al.* Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. *J. Infect. Dis.* **203**, 921–929 (2011).

188. Nakaya, H. I. *et al.* Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.* **12**, 786–795 (2011).

189. Brooks, J. P. & Lee, E. K. Analysis of the consistency of a mixed integer programming-based multi-category constrained discriminant model. *Ann. Oper. Res.* **174**, 147–168 (2010).

190. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

191. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).

192. Dudoit, S., Fridlyand, J. & Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–86 (2002).

193. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).

194. Obermoser, G. *et al.* Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity* **38**, 831–844 (2013).

195. Pulendran, B. Learning immunology from the yellow fever vaccine: Innate immunity to systems vaccinology. *Nat. Rev. Immunol.* **9**, 741–747 (2009).

196. Hagan, T. *et al.* Transcriptional atlas of the human immune response to 13 vaccines reveals a common predictor of vaccine-induced antibody responses. *bioRxiv* **10**, 2022.04.20.488939 (2022).

197. Querec, T. *et al.* Yellow fever vaccine YF-17D activates multiple dendritic cell subsets via TLR2, 7, 8, and 9 to stimulate polyvalent immunity. *J. Exp. Med.* **203**, 413–424 (2006).

198. Santos, A. P., Matos, D. C. S., Bertho, A. L., Mendonça, S. C. F. & Marcovistz, R. Detection of TH1/TH2 cytokine signatures in yellow fever 17DD first-time vaccinees through ELISpot assay. *Cytokine* **42**, 152–155 (2008).

199. Hou, J. *et al.* A Systems Vaccinology Approach Reveals Temporal Transcriptomic Changes of Immune Responses to the Yellow Fever 17D Vaccine. *J. Immunol.* **199**, 1476–1489 (2017).

200. Wang, L. T. *et al.* A Potent Anti-Malarial Human Monoclonal Antibody Targets Circumsporozoite Protein Minor Repeats and Neutralizes Sporozoites in the Liver. *Immunity* **53**, 733-744.e8 (2020).

201. Wu, R. L. *et al.* Low-Dose Subcutaneous or Intravenous Monoclonal Antibody to Prevent Malaria. *N. Engl. J. Med.* **387**, 397–407 (2022).

202. Vahey, M. T. *et al.* Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *J. Infect. Dis.* **201**, 580–589 (2010).

203. Kazmin, D. *et al.* Systems analysis of protective immune responses to RTS,S malaria vaccination in humans. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2425–2430 (2017).

204. Li, S. *et al.* Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* **169**, 862-877.e17 (2017).

205. Nakaya, H. I. *et al.* Systems biology of immunity to MF59-adjuvanted versus nonadjuvanted trivalent seasonal influenza vaccines in early childhood. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 1853–1858 (2016).

206. Filtjens, J. *et al.* Contribution of the Ly49E natural killer receptor in the immune response to Plasmodium berghei infection and control of hepatic parasite development. *PLoS One* **9**, (2014).

207. Rydyznski, C. *et al.* Generation of cellular immune memory and B-cell immunity is impaired by natural killer cells. *Nat. Commun.* **6**, (2015).

208. Levin, M. J. *et al.* Decline in Varicella-Zoster Virus (VZV)-Specific Cell-Mediated Immunity with Increasing Age and Boosting with a High-Dose VZV Vaccine. *J. Infect. Dis.* **188**, 1336–1344 (2003).

209. Patel, N. P. *et al.* Impact of Zostavax Vaccination on T-Cell Accumulation and Cutaneous Gene Expression in the Skin of Older Humans After Varicella Zoster Virus Antigen-Specific Challenge. *J. Infect. Dis.* **218**, S88–S98 (2018).

210. Qi, Q. *et al.* Defective T Memory Cell Differentiation after Varicella Zoster Vaccination in Older Individuals. *PLoS Pathog.* **12**, 1–23 (2016).

211. Coffman, R. L., Sher, A. & Seder, R. A. Vaccine adjuvants: Putting innate immunity to work. *Immunity* **33**, 492–503 (2010).

212. Galson, J. D., Trück, J., Kelly, D. F. & Van Der Most, R. Investigating the effect of AS03 adjuvant on the plasma cell repertoire following pH1N1 influenza vaccination. *Sci. Rep.* **6**, 1–11 (2016).

213. Khurana, S. *et al.* MF59 adjuvant enhances diversity and affinity of antibody-mediated immune response to pandemic influenza vaccines. *Sci. Transl. Med.* **3**, 1–10 (2011).

214. Khurana, S. *et al.* AS03-adjuvanted H5N1 vaccine promotes antibody diversity and affinity maturation, NAI titers, cross-clade H5N1 neutralization, but not H1N1 cross-subtype neutralization. *npj Vaccines* **3**, 40 (2018).

215. Givord, C. *et al.* Activation of the endoplasmic reticulum stress sensor IRE1α by the vaccine adjuvant AS03 contributes to its immunostimulatory properties. *npj Vaccines* **3**, (2018).

216. Galli, G. *et al.* Adjuvanted H5N1 vaccine induces early CD4+ T cell response that predicts long-term persistence of protective antibody levels. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3877–3882 (2009).

217. Howard, L. M. et al. Cell-based systems biology analysis of human AS03-adjuvanted H5N1 avian influenza vaccine responses: A phase i randomized controlled trial. PLoS ONE **12**, (2017).

218. Howard, L. M. *et al.* AS03-adjuvanted H5N1 avian influenza vaccine modulates early innate immune signatures in human peripheral blood mononuclear cells. *J. Infect. Dis.* **219**, 1786–1798 (2019).

219. de Mot, L. *et al.* Transcriptional profiles of adjuvanted hepatitis B vaccines display variable interindividual homogeneity but a shared core signature. *Sci. Transl. Med.* **12**, 1–14 (2020).

220. Zak, D. E. *et al.* Merck Ad5/HIV induces broad innate immune activation that predicts CD8+ T-cell responses but is attenuated by preexisting Ad5 immunity. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E3503–E3512 (2012).

221. Santini, S. M. *et al.* Interferon-α-conditioned human monocytes combine a TH1-orienting attitude with the induction of autologous TH17 responses: Role of IL-23 and IL-12. *PLoS One* **6**, (2011).

222. Volpe, E. *et al.* Multiparametric analysis of cytokine-driven human Th17 differentiation reveals a differential regulation of IL-17 and IL-22 production. *Blood* **114**, 3610–3614 (2009).

223. Fourati, S. *et al.* Pre-vaccination inflammation and B-cell signalling predict age-related hyporesponse to hepatitis B vaccination. *Nat. Commun.* **7**, 1–12 (2016).

224. Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P. & Khatri, P. Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Res.* **45**, 1–14 (2017).

225. Fourati, S. *et al.* An innate immune activation state prior to vaccination predicts responsiveness to multiple vaccines *Biorxiv* 1–22 (2021).

226. Weinberger, B. *et al.* Impaired immune response to primary but not to booster vaccination against hepatitis B in older adults. *Front. Immunol.* **9**, (2018).

227. Mulè, M. P. scglmmr. (2022). doi:10.5281/zenodo.6536393

228. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 1–14 (2016).

229. Bacher, R. *et al.* SCnorm: Robust normalization of single-cell RNA-seq data. *Nat. Methods* **14**, 584–586 (2017).

230. 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. *Nat. Commun.* **9**, (2018).

231. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).

232. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 574574 (2019).

233. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 1–15 (2019).

234. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol.* **22**, 1–20 (2021).

235. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1098 (2013).

236. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).

237. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).

238. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).

239. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biol.* **23**, 2021.07.07.451498 (2022).

240. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).

241. Govek, K. W. *et al.* Single-cell transcriptomic analysis of mIHC images via antigen mapping. *Sci. Adv.* **7**, 672501 (2021).

242. Trong, T. N. *et al.* Semisupervised Generative Autoencoder for Single-Cell Data. *J. Comput. Biol.* **27**, 1190–1203 (2020).

243. Li, B. *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).

244. Andersen, M. N., Al-Karradi, S. N. H., Kragstrup, T. W. & Hokland, M. Elimination of erroneous results in flow cytometry caused by antibody binding to Fc receptors on human monocytes and macrophages. *Cytom. Part A* **89**, 1001–1009 (2016).

245. Lun, A. T. L. *et al.* EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 1–9 (2019).

246. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, 1–10 (2020).

247. Slyper, M. *et al.* A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).

248. Buus, T. B. *et al.* Improving oligo-conjugated antibody signal in multimodal single-cell analysis. *Elife* **10**, 1–20 (2021).

249. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society: Series B* **63**, 411–423 (2001).

250. Liu, C. *et al.* Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* **184**, 1836-1857.e22 (2021).

251. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

252. Swanson, E. *et al.* Integrated single cell analysis of chromatin accessibility and cell surface markers. *bioRxiv* 2020.09.04.283887 (2020).

253. Toubal, A., Nel, I., Lotersztajn, S. & Lehuen, A. Mucosal-associated invariant T cells and disease. *Nat. Rev. Immunol.* **19**, 643–657 (2019).

254. Kjer-Nielsen, L. *et al.* MR1 presents microbial vitamin B metabolites to MAIT cells. *Nature* **491**, 717–723 (2012).

255. Pittet, M. J., Speiser, D. E., Valmori, D., Cerottini, J.-C. & Romero, P. Cutting Edge: Cytolytic Effector Function in Human Circulating CD8 + T Cells Closely Correlates with CD56 Surface Expression . *J. Immunol.* **164**, 1148–1152 (2000).

256. Van Acker, H. H., Capsomidis, A., Smits, E. L. & Van Tendeloo, V. F. CD56 in the immune system: More than a marker for cytotoxicity? *Front. Immunol.* **8**, 1–9 (2017).

257. Legoux, F. *et al.* Molecular mechanisms of lineage decisions in metabolite-specific T cells. *Nat. Immunol.* **20**, 1244–1255 (2019).

258. Salou, M. *et al.* A common transcriptomic program acquired in the thymus defines tissue residency of MAIT and NKT subsets. *J. Exp. Med.* **216**, 133–151 (2019).

259. Cheng, Z. Y., He, T. T., Gao, X. M., Zhao, Y. & Wang, J. ZBTB Transcription Factors: Key Regulators of the Development, Differentiation and Effector Function of T Cells. *Front. Immunol.* **12**, 1–19 (2021).

260. Raberger, J. *et al.* The transcriptional regulator PLZF induces the development of CD44 high memory phenotype T cells. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 17919–17924 (2008).

261. Park, D. *et al.* Differences in the molecular signatures of mucosal-associated invariant T cells and conventional T cells. *Sci. Rep.* **9**, 1–10 (2019).

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

263. Lian, Q. *et al.* Artificial-cell-type aware cell-type classification in CITE-seq. *Bioinformatics* **36**, I542–I550 (2020).

264. Wang, X. *et al.* BREM-SC: A bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic Acids Res.* **48**, 5814–5824 (2020).

265. Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics* **36**, 4137–4143 (2020).

266. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39**, 813–818 (2021).

267. Roelli, P., bbimber, Flynn, B., santiagorevale & Gui, G. Hoohm/CITE-seq-Count: 1.4.2. (2019). doi:10.5281/ZENODO.2590196

268. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

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

270. Fraley, C., Raftery, A. E., Murphy, T. B. & Scrucca, L. MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. (2012).

271. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *Arxiv* (2018).

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

273. Waltman, L. & Van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, (2013).

274. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. *bioRxiv* (2019). doi:10.1101/060012

275. Liston, A., Humblet-Baron, S., Duffy, D. & Goris, A. Human immune diversity: from evolution to modernity. *Nat. Immunol.* **22**, 1479–1489 (2021).

276. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.* **10**, 116–125 (2009).

277. Moncunill, G. *et al.* Transcriptional correlates of malaria in RTS,S/AS01-vaccinated African children: A matched case-control study. *Elife* **11**, 1–30 (2022).

278. Tsang, J. S. *et al.* Improving Vaccine-Induced Immunity: Can Baseline Predict Outcome? *Trends Immunol.* **41**, 457–465 (2020).

279. Xhonneux, L. P. *et al.* Transcriptional networks in at-risk individuals identify signatures of type 1 diabetes progression. *Sci. Transl. Med.* **13**, 1–16 (2021).

280. Germain, R. N. & Schwartzberg, P. L. The human condition: An immunological perspective. *Nat. Immunol.* **12**, 369–372 (2011).

281. Roederer, M. *et al.* The genetic architecture of the human immune system: A bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).

282. Garçon, N., Vaughn, D. W. & Didierlaurent, A. M. Development and evaluation of AS03, an Adjuvant System containing α-tocopherol and squalene in an oil-in-water emulsion. *Expert Rev. Vaccines* **11**, 349–366 (2012).

283. Ellebedy, A. H. *et al.* Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat. Immunol.* **17**, 1226–1234 (2016).

284. Chang, W. C. *et al.* Genetic variants of PPAR-gamma coactivator 1B augment NLRP3-mediated inflammation in gouty arthritis. *Rheumatol. (United Kingdom)* **56**, 457–466 (2017).

285. Segovia, M. *et al.* Targeting TMEM176B Enhances Antitumor Immunity and Augments the Efficacy of Immune Checkpoint Blockers by Unleashing Inflammasome Activation. *Cancer Cell* **35**, 767-781.e6 (2019).

286. Nordmann, A., Wixler, L., Boergeling, Y., Wixler, V. & Ludwig, S. A new splice variant of the human guanylate-binding protein 3 mediates anti-influenza activity through inhibition of viral transcription and replication. *FASEB J.* **26**, 1290–1300 (2012).

287. Hsiang, T.-Y., Zhao, C. & Krug, R. M. Interferon-Induced ISG15 Conjugation Inhibits Influenza A Virus Gene Expression and Replication in Human Cells. *J. Virol.* **83**, 5971–5977 (2009).

288. Li, Y. *et al.* Activation of RNase L is dependent on OAS3 expression during infection with diverse human viruses. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2241–2246 (2016).

289. Qu, H. *et al.* Influenza A Virus-induced expression of ISG20 inhibits viral replication by interacting with nucleoprotein. *Virus Genes* **52**, 759–767 (2016).

290. Fantuzzi, L. *et al.* Loss of CCR2 expression and functional response to monocyte chemotactic protein (MCP-1) during the differentiation of human monocytes: Role of secreted MCP-1 in the regulation of the chemotactic response. *Blood* **94**, 875–883 (1999).

291. Kuss-Duerkop, S. K. et al. Influenza virus differentially activates mTORC1 and mTORC2 signaling to maximize late stage replication. PLoS Pathogens **13**, (2017).

292. Weichhart, T. *et al.* The TSC-mTOR Signaling Pathway Regulates the Innate Inflammatory Response. *Immunity* **29**, 565–577 (2008).

293. Kumar, V. *et al.* mTOR/HIF1α-mediated aerobic glycolysis as metabolic basis for trained immunity. *Science (80-.).* **345**, 1–18 (2014).

294. Marçais, A. *et al.* The metabolic checkpoint kinase mTOR is essential for IL-15 signaling during the development and activation of NK cells. *Nat. Immunol.* **15**, 749–757 (2014).

295. Mao, Q., Wang, L., Tsang, I. W. & Sun, Y. Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2227–2241 (2017).

296. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).

297. Chari, T., Banerjee, J. & Pachter, L. The Specious Art of Single-Cell Genomics. *BioRxiv* 1–25 (2021).

298. Alquicira-Hernandez, J., Powell, J. E. & Phan, T. G. No evidence that plasmablasts transdifferentiate into developing neutrophils in severe COVID-19 disease. *Clin. Transl. Immunol.* **10**, 1–7 (2021).

299. Bloes, D. A., Kretschmer, D. & Peschel, A. Enemy attraction: Bacterial agonists for leukocyte chemotaxis receptors. *Nat. Rev. Microbiol.* **13**, 95–104 (2015).

300. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science (80-.).* **339**, 786–91 (2013).

301. Sallusto, F. *et al.* Rapid and coordinated switch in chemokine receptor expression during dendritic cell maturation. *Eur. J. Immunol.* **28**, 2760–2769 (1998).

302. Chhatbar, C. & Prinz, M. The roles of microglia in viral encephalitis: from sensome to therapeutic targeting. *Cell. Mol. Immunol.* **18**, 250–258 (2021).

303. Hoshino, K. *et al.* Toll-Like Receptor 4 (TLR4)-Deficient Mice Are Hyporesponsive to Lipopolysaccharide: Evidence for TLR4 as the Lps Gene Product. *J. Immunol.* **162**, 3749–3752 (1999).

304. Wensveen, F. M. *et al.* Apoptosis threshold set by noxa and Mcl-1 after T cell activation regulates competitive selection of high-affinity clones. *Immunity* **32**, 754–765 (2010).

305. Wensveen, F. M. *et al.* BH3-only protein Noxa regulates apoptosis in activated B cells and controls high-affinity antibody formation. *Blood* **119**, 1440–1449 (2012).

306. Gricks, C. S. *et al.* Differential regulation of gene expression following CD40 activation of leukemic compared to healthy B cells. *Blood* **104**, 4002–4009 (2004).

307. Shimabukuro-Vornhagen, A. *et al.* Inhibition of Protein Geranylgeranylation Specifically Interferes with CD40-Dependent B Cell Activation, Resulting in a Reduced Capacity To Induce T Cell Immunity. *J. Immunol.* **193**, 5294–5305 (2014).

308. Billerbeck, E. *et al.* Analysis of CD161 expression on human CD8+ T cells defines a distinct functional subset with tissue-homing properties. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3006–3011 (2010).

309. Tastan, C. *et al.* Tuning of human MAIT cell activation by commensal bacteria species and MR1-dependent T-cell presentation. *Mucosal Immunol.* **11**, 1591–1605 (2018).

310. Wilgenburg, B. van *et al.* MAIT cells contribute to protection against lethal influenza infection in vivo. *Nat. Commun.* **9**, (2018).

311. Ussher, J. E. *et al.* CD161++CD8+ T cells, including the MAIT cell subset, are specifically activated by IL-12+IL-18 in a TCR-independent manner. *Eur. J. Immunol.* **44**, 195–203 (2014).

312. Hogan, M. J. & Pardi, N. MRNA Vaccines in the COVID-19 Pandemic and Beyond. *Annu. Rev. Med.* **73**, 17–39 (2022).

313. Arunachalam, P. S. *et al.* Systems vaccinology of the BNT162b2 mRNA vaccine in humans. *Nature* **596**, 410–416 (2021).

314. Pardi, N. *et al.* Nucleoside-modified mRNA vaccines induce potent T follicular helper and germinal center B cell responses. *J. Exp. Med.* **215**, 1571–1588 (2018).

315. Liang, F. *et al.* Efficient Targeting and Activation of Antigen-Presenting Cells In Vivo after Modified mRNA Vaccine Administration in Rhesus Macaques. *Mol. Ther.* **25**, 2635–2647 (2017).

316. Farmer, R., Apps, R. & Tsang, J. S. HDStIM: High Dimensional Stimulation Immune Mapping. (2022).

317. Doyle, S. E. *et al.* IRF3 Mediates a TLR3/TLR4-Specific Antiviral Gene Program. *Immunity* **17**, 251–263 (2002).

318. Turner, J. S. *et al.* Human germinal centres engage memory and naive B cells after influenza vaccination. *Nature* **586**, 127–132 (2020).

319. Gupta, N. T. *et al.* Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).

320. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39**, 451–461 (2021).

321. Oh, J. Z. *et al.* TLR5-mediated sensing of gut microbiota is necessary for antibody responses to seasonal influenza vaccination. *Immunity* **41**, 478–492 (2014).

322. Wimmers, F. *et al.* The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination. *Cell* **184**, 3915-3935.e21 (2021).

323. Keenan, A. B. *et al.* ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.* **47**, W212–W224 (2019).

324. McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).

325. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

326. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **5**, (2016).

327. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. *BioRxiv* 1–29 (2019). doi:10.1101/060012

328. Ribas, A. & Wolchok, J. D. Cancer immunotherapy using checkpoint blockade. *Science (80-.).* **359**, 1350–1355 (2018).

329. Xu, C. *et al.* Comparative safety of immune checkpoint inhibitors in cancer: systematic review and network meta-analysis. *BMJ* **363**, k4226 (2018).

330. Dougan, M. & Pietropaolo, M. Time to dissect the autoimmune etiology of cancer antibody immunotherapy. *J Clin Invest* **130**, 51–61 (2020).

331. Martins, F. *et al.* Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nat Rev Clin Oncol* **16**, 563–580 (2019).

332. Puzanov, I. *et al.* Managing toxicities associated with immune checkpoint inhibitors: consensus recommendations from the Society for Immunotherapy of Cancer (SITC) Toxicity Management Working Group. *J Immunother Cancer* **5**, 95 (2017).

333. Brahmer, J. R. *et al.* Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol* **36**, 1714–1768 (2018).

334. Haanen, J. *et al.* Rechallenge patients with immune checkpoint inhibitors following severe immune-related adverse events: review of the literature and suggested prophylactic strategy. *J Immunother Cancer* **8**, (2020).

335. Ramos-Casals, M. *et al.* Immune-related adverse events of checkpoint inhibitors. *Nat. Rev. Dis. Prim.* **6**, 38 (2020).

336. Pauken, K. E., Dougan, M., Rose, N. R., Lichtman, A. H. & Sharpe, A. H. Adverse Events Following Cancer Immunotherapy: Obstacles and Opportunities. *Trends Immunol* **40**, 511–523 (2019).

337. Kuehn, H. S. *et al.* Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. *Science (80-.).* **345**, 1623–1627 (2014).

338. Klocke, K., Sakaguchi, S., Holmdahl, R. & Wing, K. Induction of autoimmune disease by deletion of CTLA-4 in mice in adulthood. *Proc Natl Acad Sci U S A* **113**, E2383-92 (2016).

339. Okazaki, T., Chikuma, S., Iwai, Y., Fagarasan, S. & Honjo, T. A rheostat for immune responses: the unique properties of PD-1 and their advantages for clinical application. *Nat Immunol* **14**, 1212–1218 (2013).

340. Abdel-Wahab, N., Shah, M., Lopez-Olivo, M. A. & Suarez-Almazor, M. E. Use of Immune Checkpoint Inhibitors in the Treatment of Patients With Cancer and Preexisting Autoimmune Disease: A Systematic Review. *Ann Intern Med* **168**, 121–130 (2018).

341. Larkin, J. *et al.* Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *N Engl J Med* **373**, 23–34 (2015).

342. Van Der Vlist, M., Kuball, J., Radstake, T. R. D. & Meyaard, L. Immune checkpoints and rheumatic diseases: What can cancer immunotherapy teach us? *Nat. Rev. Rheumatol.* **12**, 593–604 (2016).

343. Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).

344. Wong, H. S. *et al.* A local regulatory T cell feedback circuit maintains immune homeostasis by pruning self-activated T cells. *Cell* **184**, 3981-3997 e22 (2021).

345. Luoma, A. M. *et al.* Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy. *Cell* **182**, 655-671.e22 (2020).

346. Rajan, A. & Zhao, C. Deciphering the biology of thymic epithelial tumors. *Mediastinum* **3**, (2019).

347. Zhao, C. & Rajan, A. Immune checkpoint inhibitors for treatment of thymic epithelial tumors: how to maximize benefit and optimize risk? *Mediastinum* **3**, (2019).

348. Krieg, C. *et al.* High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy. *Nat. Med.* **24**, 144–153 (2018).

349. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013 e20 (2018).

350. Wu, T. D. *et al.* Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature* **579**, 274–278 (2020).

351. Caushi, J. X. *et al.* Transcriptional programs of neoantigen-specific TIL in anti-PD-1-treated lung cancers. *Nature* **596**, 126–132 (2021).

352. Nehar-Belaid, D. *et al.* Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat Immunol* **21**, 1094–1106 (2020).

353. Pinal-Fernandez, I. *et al.* Identification of distinctive interferon gene signatures in different types of myositis. *Neurology* **93**, e1193–e1204 (2019).

354. Binnewies, M. *et al.* Unleashing Type-2 Dendritic Cells to Drive Protective Antitumor CD4(+) T Cell Immunity. *Cell* **177**, 556-571 e16 (2019).

355. Sabatini, D. M. mTOR and cancer: insights into a complex relationship. *Nat Rev Cancer* **6**, 729–734 (2006).

356. Perl, A. Activation of mTOR (mechanistic target of rapamycin) in rheumatic diseases. *Nat Rev Rheumatol* **12**, 169–182 (2016).

357. Simoni, Y. *et al.* Bystander CD8(+) T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature* **557**, 575–579 (2018).

358. Albrecht, I. *et al.* Persistence of effector memory Th1 cells is regulated by Hopx. *Eur J Immunol* **40**, 2993–3006 (2010).

359. Weng, N. P., Araki, Y. & Subedi, K. The molecular basis of the memory T cell response: differential gene expression and its epigenetic regulation. *Nat Rev Immunol* **12**, 306–315 (2012).

360. Omilusik, K. D. *et al.* Transcriptional repressor ZEB2 promotes terminal differentiation of CD8+ effector and memory T cell populations during infection. *J Exp Med* **212**, 2027–2039 (2015).

361. Cook, M. E., Jarjour, N. N., Lin, C. C. & Edelson, B. T. Transcription Factor Bhlhe40 in Immunity and Autoimmunity. *Trends Immunol* **41**, 1023–1036 (2020).

362. Eltzschig, H. K. & Carmeliet, P. Hypoxia and inflammation. *N Engl J Med* **364**, 656–665 (2011).

363. Esfahani, K. *et al.* Targeting the mTOR pathway uncouples the efficacy and toxicity of PD-1 blockade in renal transplantation. *Nat. Commun.* **10**, 1–9 (2019).

364. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).

365. Lenth, R. emmeans: Estimated Marginal Means, aka Least-Squares Means, https://CRAN.R-project.org/package=emmeans. (2019).

366. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

367. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498–D503 (2020).

368. Dolgalev, I. msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. R package version 6.2.1. https://CRAN.R-project.org/package=msigdbr. (2018).

369. Shahabi, V. *et al.* Gene expression profiling of whole blood in ipilimumab-treated patients for identification of potential biomarkers of immune-related gastrointestinal adverse events. *J. Transl. Med.* **11**, 1–11 (2013).

370. Yao, C. *et al.* Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. *Nat. Immunol.* **20**, 890–901 (2019).

371. Shen, L. GeneOverlap: Test and visualize gene overlaps. R package version 1.18.0. https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html. (2018). Available at: https://bioconductor.org/packages/release/bioc/html/GeneOverlap.html.

372. Jardine, L. et al. Blood and immune development in human fetal bone marrow and Down syndrome. Nature **598**, (2021).

373. Lee, J. C. *et al.* Gene expression profiling of CD8 + T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J. Clin. Invest.* **121**, 4170–4179 (2011).

374. Locci, M. *et al.* Human circulating PD-1+CXCR3-CXCR5+ memory Tfh cells are highly functional and correlate with broadly neutralizing HIV antibody responses. *Immunity* **39**, 758–769 (2013).

6 APPENDIX – ADDITIONAL MATERIALS FOR CHAPTER 2

Supplementary Materials

Normalizing and denoising protein expression data from droplet-based single cell profiling

Matthew P. Mulè[1,3,4] Andrew J. Martins[1,4] and John S. Tsang[1,2]
1. Multiscale Systems Biology Section, Laboratory of Immune System Biology, National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH)
2. NIH Center for Human Immunology (CHI), National Institutes of Health (NIH)
3. NIH-Oxford-Cambridge Scholars Program, Department of Medicine, Cambridge University
4. These authors contributed equally Matthew P. Mulè, Andrew J. Martins
Correspondence to: john.tsang@nih.gov

Contents:

The method presented in this paper is available as an open source R package "dsb" available on CRAN. For up to date tutorials, please see the package documentation and vignettes:

https://CRAN.R-project.org/package=dsb

Supplementary Note

Robustness of protein-specific noise estimation assessed by using different approaches to define empty/background droplets

The dsb package utilizes the raw (unfiltered) output of UMI count aligners such as Cell Ranger, Kallisto[1] or as we used here, CITE-seq Count[2]. The unfiltered output (for example, in Cell Ranger, the *raw* output) of droplet barcodes versus UMI counts includes all cell containing and empty (or "background") droplets, both of which can be inferred using thresholding methods based on the mRNA and protein library sizes in combination with algorithms like EmptyDrops[3] to distinguish cells from background noise (as done by default by Cell Ranger)–see dsb package documentation tutorial. In all datasets analyzed by us to date, a considerable number (at least 50,000 after QC) of background droplets (i.e., barcodes inferred to not contain at least one cell) can be found using library size based thresholding (see below for robustness assessments). The protein counts derived from these background droplets reflect contributions from ambient antibodies, which as shown in the main text, were highly correlated with the protein counts detected in unstained control cells included in our experiment. Thus, as discussed in the main text, protein counts in empty droplets can serve as an estimate of the expected ambient levels of antibodies. To assess the robustness of estimating protein-specific noise in relation to how background droplets are defined, we compared three approaches to define background droplets. As detailed in our previous report[4], due to the number of samples included in our experiment, demultiplexing samples required data from both sample barcode ("cell hashing") antibodies and mRNA (for genetic based demultiplexing, i.e., by cross referencing independently generated patient genotype data using demuxlet, see Methods and Kotliarov *et. al.* 2020). After removing doublets and defining singlets on the basis of data from both the hashing antibodies and genotypes, the remaining (non-doublet, non-singlet) droplets were used to define background droplets in two different ways. First, "Library size background droplets" were defined solely based on library size information where we used clear breaks in the distribution of protein library sizes across the remaining droplets followed by removal of droplets in the top 10^{th} percentile based on the mRNA library size in order to eliminate droplets containing low quality cells. The library size approach to define background droplets is most compatible with experiments that do not have sample multiplexing or hashing antibody data, such as the external CITE-seq datasets from 10X Genomics used in this paper (Fig. 3 and Supplementary Fig. 4).

The second background droplet inference method we tested requires CITE-seq experimental workflows similar to ours, where many samples are multiplexed in the same experiment using sample barcoding antibodies (and/or genetic based demultiplexing). After using Seurat's K-medoids function to computationally classify cell barcodes as containing singlets, doublets, or negatives based on the hashing antibody counts, we defined "Hashing background droplets" as those classified as "negative" by this demultiplexing software. These droplets had staining below the threshold to be called positive for any one of the hashing antibodies and

therefore in principle, their antibody counts should reflect only ambient capture. Such hashing "negatives" were an order of magnitude fewer in number than those determined by library size above, largely due to the threshold used for determining whether a droplet is included in the hash demultiplexing pipeline (the top 35,000 barcodes from each lane). Hashing background droplets were further filtered to: 1) include only droplets classified as "ambiguous" by SNP demultiplexing (via demuxlet), i.e., these cannot be attributed to a single or multiple distinct donors based on cross-referencing mRNA reads in the droplet with independently generated genotype data, and 2) exclude any droplet with >80 unique mRNAs to remove cell-containing droplets with low-quality mRNA capture. Using this alternative method to define background droplets, we similarly observed that the relative amount of antibody was highly correlated between unstained cells and these background droplets (along the unity line in Supplementary Fig. 1b, top).

Interestingly, while the correlation was similarly high, antibody levels in unstained cells or in hashing background droplets were greater than those in library size background droplets (Supplementary Fig. 1b top vs bottom). The greater magnitude of antibody counts by a multiplicative factor in log-count space (slope in bottom panel of Supplementary Fig. 1b is 1.24 with near zero intercept) suggests that unstained cells and demultiplexing background droplets capture additional antibodies. Unstained cells may serve as an additional antibody capturing "reservoir", e.g., due to non-specific (or specific) binding of the ambient antibody remaining after multiple wash steps. However, this would not explain their concordance with demultiplexing background, which, as supported by both genetic (via demuxlet classification) and barcoding antibody (via Seurat k-medoids classification) data, should have a low chance of containing fully intact cells. It is still possible, despite filtering out droplets with low mRNA counts, that demultiplexing background droplets contained some very low-quality cells or cell membrane debris that together could capture additional antibodies from the environment via specific/non-specific binding. Demultiplexing background droplets could also have more ambient mRNA (as described above in order to be included in the hashing antibody demultiplexing step) than droplets defined using the protein library size distribution alone, and thus they (as also in the unstained control cell droplets) could conceivably serve as an additional set of free antibody-capturing molecules. Importantly, however, we emphasize the difference between empty/background droplets defined using protein library size distribution versus hashing antibody demultiplexing had negligible effect in the resulting dsb normalized values (see below).

We further investigated a third approach to estimate protein background noise–the mean of each protein across the subset of *stained cells* that were inferred to belong to the "*negative"* population for each protein. Without dsb rescaling, we fit a two component Gaussian mixture model to the log + 1 transformed count of each protein across single cells, resulting in 2 populations of cells: those positive or negative for the protein. Each protein's background mean, "*A*" (see Supplementary Fig.1a), reflects the average log transformed count of the non-staining cell population for that protein, i.e., cells that do not express that protein. The protein level in

unstained controls and empty drops were both highly correlated with *A* (Supplementary Fig. 1c). Thus, antibody levels in unstained droplets on average are similar to those in droplets with stained cells not expressing the target protein.

We thus have tested three different ways to estimate the average background protein noise correlated across droplets. We further found that the noise signal captured in library size background droplets appears to be universally found in data from all of the droplet-based oligo barcoded antibody experiments we examined and is thus a generalizable method of estimating noise.

Importance of using isotype control antibodies for estimating cell-intrinsic normalization factors

Our method is compatible with experiments lacking isotype controls by either not removing the cell-specific technical variation (use *denoise.counts = FALSE* in dsb*)* or by removing the technical component with a single fitted parameter, the per-cell mean of the background protein population (parameters *denoise.counts = TRUE, use.isotype.control = FALSE*). However, additional analyses further support our findings that inclusion of isotype controls benefits cell to cell technical noise correction (step II). Despite the ability of μ1 alone to provide information about the cell-intrinsic technical component, we recommend the inclusion of multiple isotype controls in CITE-seq experiments to serve as anchors for better estimation of technical normalization factors because μ1 alone may carry signals beyond those from technical factors (e.g., low-level antigen specific binding). In our data for example, μ1 exhibited greater correlation with μ2 than did the mean of the isotype controls (Supplementary Figs. 3b,c), including when sub-sampling random draws of four proteins from those used to compute μ1 within each cell to assess whether signal from four background proteins is equivalent to that of four isotype controls (Supplementary Fig. 3d). Furthermore, even with isotype controls as anchors, the estimated cell-intrinsic background may encompass signals from non-specific binding to surface Fc receptors. Cell types such as monocytes with higher relative Fc receptor expression may thus receive more correction than other cell types. However, empirically we have not found this to have adverse effects on normalized values in populations such as monocytes, cell type identification, or downstream analysis (Supplementary Figs. 6b, c). Careful blocking of Fc receptors before antibody staining, which is standard practice and was performed in our experiments, likely contributed to mitigating this effect.

Robustness of dsb normalized values to background droplet definition

Given the strong correlation observed between average protein levels in unstained control cells and both empty drops and droplets with stained cells expressing background level of the protein (Supplementary Fig. 1b,c), ambient antibodies appear to capture the major noise component that

contributes to each protein's specific noise floor. In external 10X Genomics CITE-seq datasets, we distinguished between empty droplets and cells using the Cell Ranger alignment tool which uses a method inspired by the EmptyDrops[3] algorithm to identify cell-containing droplets. The number of estimated cell-containing droplets depends on the number of cells loaded during droplet generation which should then inform the value input to the Cell Ranger *expect_cells* parameter; typical experiments recover on the order of 10^3 to10^4 droplets. Empty droplets capturing ambient ADTs, typically between $5x10^4$ to $1x10^5$ in number, can be robustly defined from the remaining, non-cell-containing barcodes in the raw output matrix. The raw output matrix lists all possible cell barcode combinations (more than 6 million barcodes in the Version 3 and Next Gem assays), many of which have no evidence of capture in the experiment (i.e. no data for mRNA or ADT reads) and empty droplets capturing ambient ADT must be subset from this output in order to avoid biasing the background estimates. The steps to complete this process are completed in a few lines of code as detailed in the dsb package documentation. A substantial subset of the cell barcodes estimated by Cell Ranger to not contain a cell had ADT reads with an order of magnitude lower protein library size compared to the cell-containing droplets. We then applied quality-control thresholds determined based on protein and mRNA counts for each dataset, for example, excluding certain "empty" droplets from being used in the background distribution that likely corresponded to potentially low-quality cells (e.g., removing empty droplets with more than 80 unique mRNA). This procedure revealed a clear population of more than 50,000 background droplets in each dataset. In some external datasets, there were two distinct background populations based on protein library size (Supplementary Fig. 7a). dsb normalized values were robust to using different background subpopulations (Supplementary Fig. 7a,b). When only the lower ADT background peak was used to simulate an experiment with extremely low background, dsb normalized values still separated canonical cell populations but were less zero-centered due to the low estimated background for some proteins (third row, Supplementary Fig. 7b). We have not encountered a dataset like this simulation scenario to date, however, in the future as antibody panels continue to increase in size, some antibodies may be titrated down to extremely low concentrations. Theoretically, this could decrease background levels in empty droplets for certain proteins to a level that could impact the first step of dsb as shown above. Our method could be easily adjusted in this hypothetical case by modifying the standardization step to accommodate lower background dispersion.

Within batch normalization vs. pooled normalization across multiple batches

The experimental design of the main dataset used here to develop our approach include n=20 unique donors distributed over two experimental batches; this presented multiple options for dsb normalization. Background/empty drops could be defined with either of the two methods described above (demultiplexing or library size), and cells could then be normalized by combining all cells / background into a single matrix and normalizing both batches together, or each batch of cells could be normalized separately, using only the empty droplets within each batch. To test how

robust the resulting dsb normalized values were to single vs multi-batch normalization, as well as to further validate the findings described above on the robustness of dsb normalized values to different definitions of background, we tested the 4 possible normalization schemes with background droplets classified by either protein library size distribution or demultiplexing, then normalized with dsb by either merging cells and background from both batches together, or normalizing each batch separately. The resulting dsb normalized values were consistently similar across all four of these normalization schemes (Supplementary Fig. 8). Since we expect ambient antibody to be a major contributor of correlated noise across cells, experimental standardization of staining time and the number of washing steps prior to droplet generation as well as use of the same pool of manually concentrated antibody on each batch could be important contributing factors in mitigating batch to batch variations. Our method is not designed as a batch effect removal tool, however, as enabled by the standard, normalized expression value scale from dsb, the approach of applying a uniform background cutoff threshold across proteins in diverse datasets can potentially help mitigate batch effects. The performance of existing batch correction tools[5,6] including single cell integration methods[7–10] on ADT data could be an area of further investigation to compare upstream dsb to other normalization methods as more datasets become available.

References

1. Melsted, P. et al. Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* **39** 813–818 (2021).
2. Roelli, P., bbimber, Flynn, B., santiagorevale & Gui, G. Hoohm/CITE-seq-Count: 1.4.2. (2019). doi:10.5281/ZENODO.2590196
3. Lun, A. T. L. *et al.* EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 1–9 (2019).
4. 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. *Nat. Med.* **26**, 618–629 (2020).
5. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
6. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
7. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
8. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
9. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
10. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods*. 18, 272–282 (2021).

Supplementary Figures

Supplementary Fig. 1. Robustness assessment of estimating ambient ADT noise in cell-containing droplets using ADT levels in empty droplets via comparison with unstained controls.

Supplementary Fig. 2. Robustness assessment of models fitted to each cell in dsb (step II part I).

Supplementary Fig. 3. Analysis of isotype control contribution to dsb technical component and comparison of dsb normalized values to centered log ratio normalization.

Supplementary Fig. 4. Analysis of dsb normalization on external CITE-seq datasets.

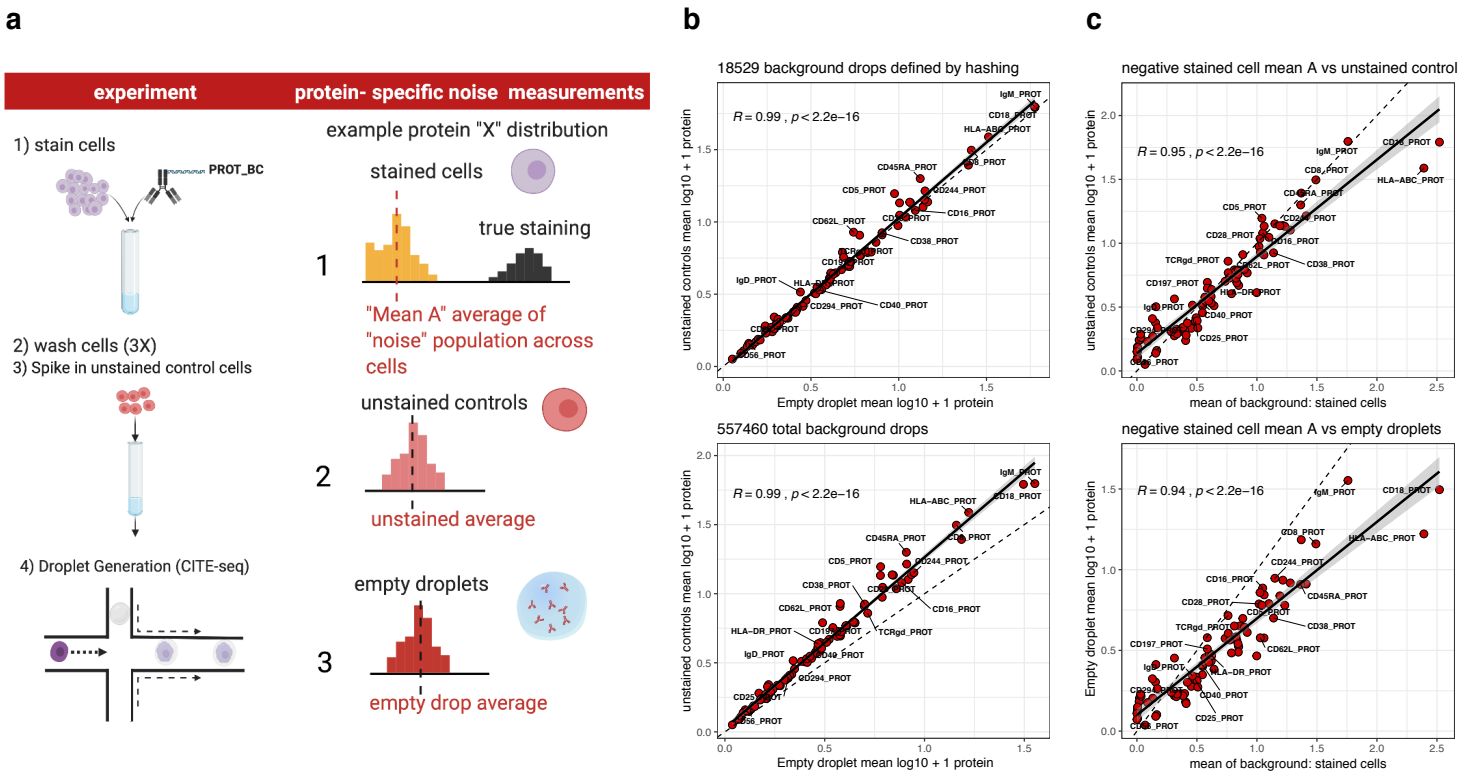
Supplementary Fig. 5. Analysis of dsb normalization on TEA-seq, ASAP-seq and Mission Bio datasets.

Supplementary Fig. 6. Manual and automatic cell type identification with protein levels after dsb normalization from healthy donor PBMC data (data from Kotliarov *et. al.* 2020).

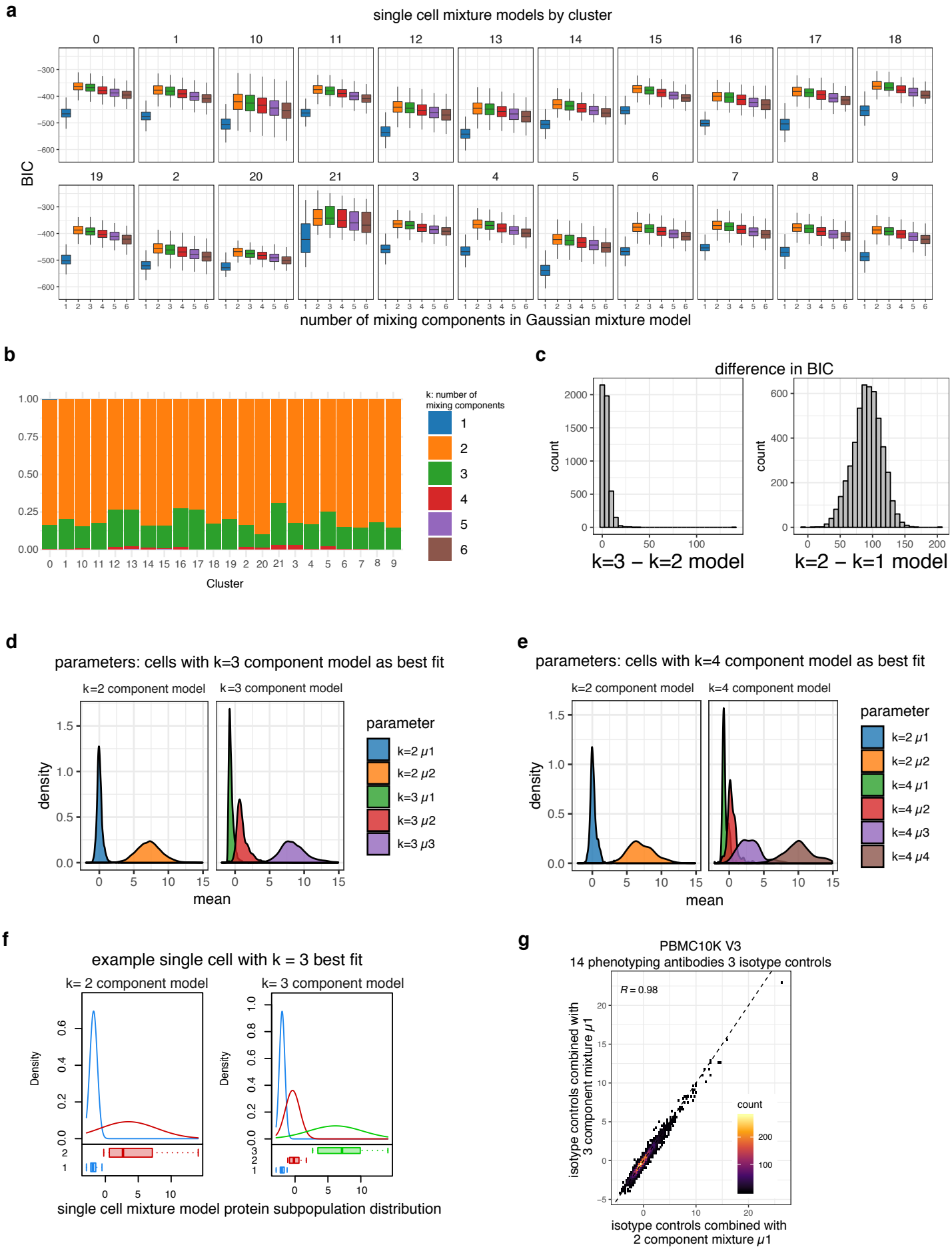
Supplementary Fig. 7. Robustness assessment of dsb normalized values to different subsets of empty droplets used for background correction with dsb.

Supplementary Fig. 8. Batch processing with dsb: analysis of merging multiple batches then normalizing, vs. separate normalization applied within each batch.

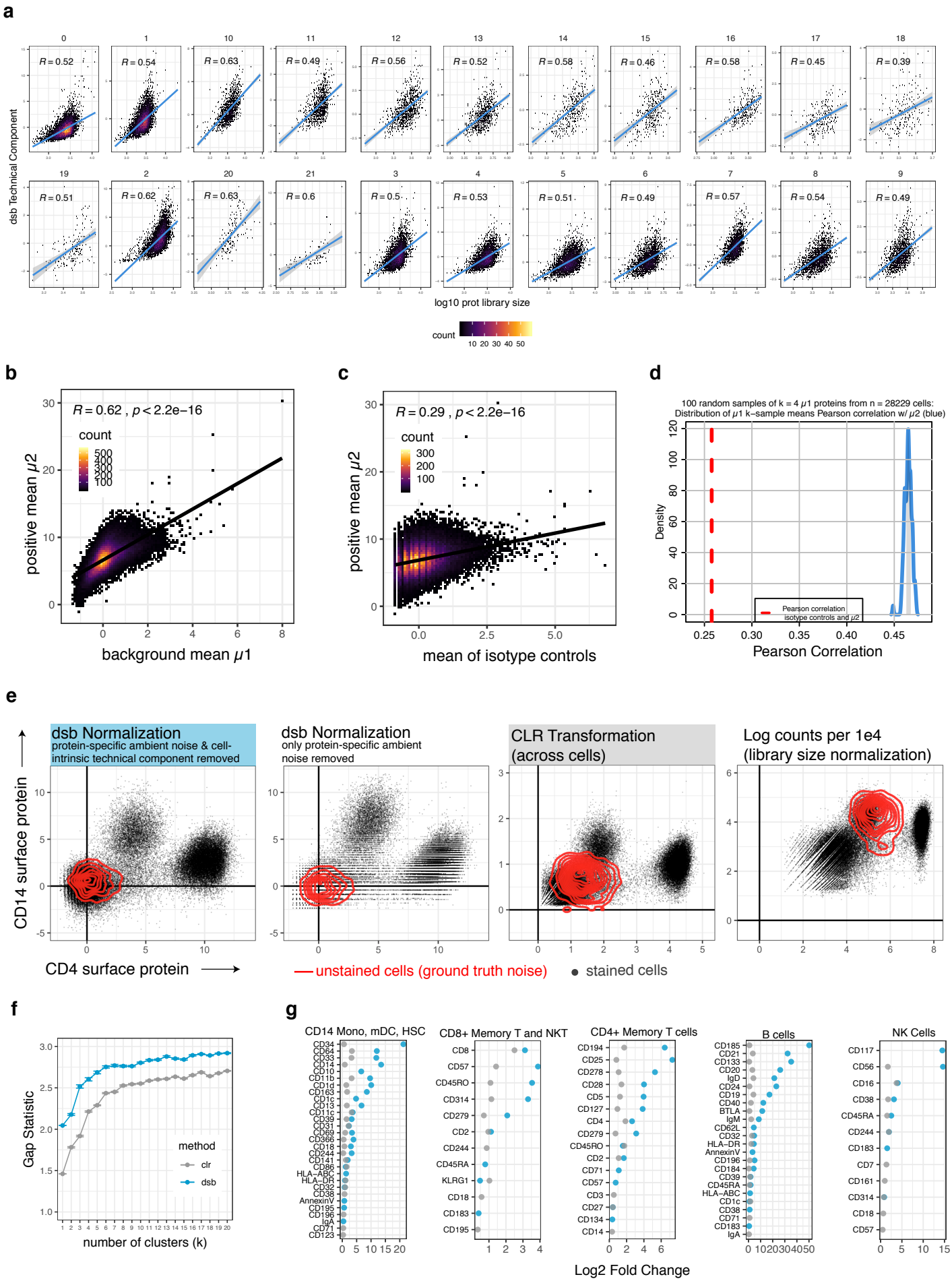
Supplementary Fig. 9. Additional figures from analysis of TEA-seq data (data from Swanson *et. al.* 2021).


a
experiment
protein- specific noise measurements
1) stain cells
PROT_BC
2) wash cells (3X)
3) Spike in unstained control cells
4) Droplet Generation (CITE-seq)
example protein "X" distribution
stained cells
true staining
1
"Mean A" average of "noise" population across cells
unstained controls
2
unstained average
empty droplets
3
empty drop average
b
18529 background drops defined by hashing
R = 0.99 , p < 2.2e−16
unstained controls mean log10 + 1 protein
Empty droplet mean log10 + 1 protein
557460 total background drops
R = 0.99 , p < 2.2e−16
unstained controls mean log10 + 1 protein
Empty droplet mean log10 + 1 protein
c
negative stained cell mean A vs unstained control
R = 0.95 , p < 2.2e−16
unstained controls mean log10 + 1 protein
mean of background: stained cells
negative stained cell mean A vs empty droplets
R = 0.94 , p < 2.2e−16
Empty droplet mean log10 + 1 protein
mean of background: stained cells

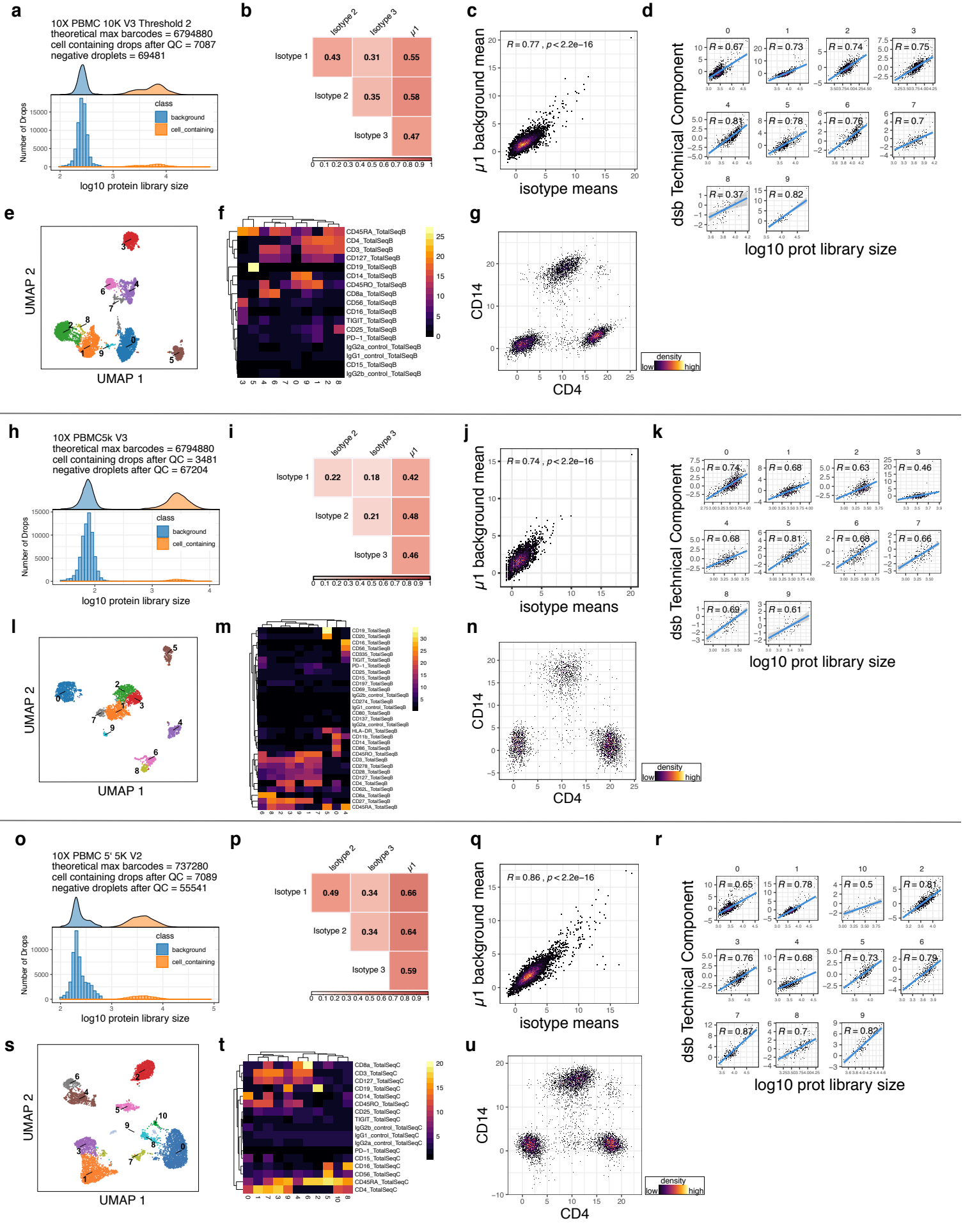
Supplementary Figure 1
a. Expanded from Fig. 1a: to assess the relative contribution of the ambient antibody component of noise correlated across droplets, three different measurements of protein-specific background noise were defined for each protein: 1) (top row, right column) for each protein, the average log transformed value of the subset of stained cells that were not part of the proteins "positive" population and comprised the "non-staining" population of cells (the negative cell population for each protein was inferred through a Gaussian mixture model fit separately to each protein, see Methods) 2) (middle row, right column): unstained control cells spiked into the cell mixture prior to droplet generation as shown in the experiment diagram (left column), 3) (bottom row, right column): empty droplets as defined by either the protein library size distribution or inferred by sample barcode antibody demultiplexing (see Methods). **b.** Pearson correlation coefficient and p value (two sided) between unstained control cells (y-axis) and empty droplets (x-axis) with empty droplets defined by either demultiplexing (top "hashing background droplets") or library size distribution (bottom, "library size background droplets", see supplemental note) **c.** Pearson correlation coefficient and p value (two sided) between y-axis: unstained controls (top panel) or library size background droplets (bottom panel) versus x-axis: the mean of the protein in stained cells that were negative for the protein ("mean A" as shown in top panel of a). In all plots the dashed line at unity (y = x) is shown for reference and the solid line is the fitted regression line with the shaded region representing the 95% confidence interval of the linear model fit centered around the fitted values. Illustration created with BioRender.com.


a
single cell mixture models by cluster
BIC
number of mixing components in Gaussian mixture model
b
k: number of mixing components
Cluster
c
difference in BIC
count
k=3 – k=2 model
k=2 – k=1 model
d
parameters: cells with k=3 component model as best fit
k=2 component model
k=3 component model
density
mean
parameter
k=2 μ1
k=2 μ2
k=3 μ1
k=3 μ2
k=3 μ3
e
parameters: cells with k=4 component model as best fit
k=2 component model
k=4 component model
density
mean
parameter
k=2 μ1
k=2 μ2
k=4 μ1
k=4 μ2
k=4 μ3
k=4 μ4
f
example single cell with k = 3 best fit
k= 2 component model
k= 3 component model
Density
single cell mixture model protein subpopulation distribution
g
PBMC10K V3
14 phenotyping antibodies 3 isotype controls
R = 0.98
isotype controls combined with 3 component mixture μ1
isotype controls combined with 2 component mixture μ1
count

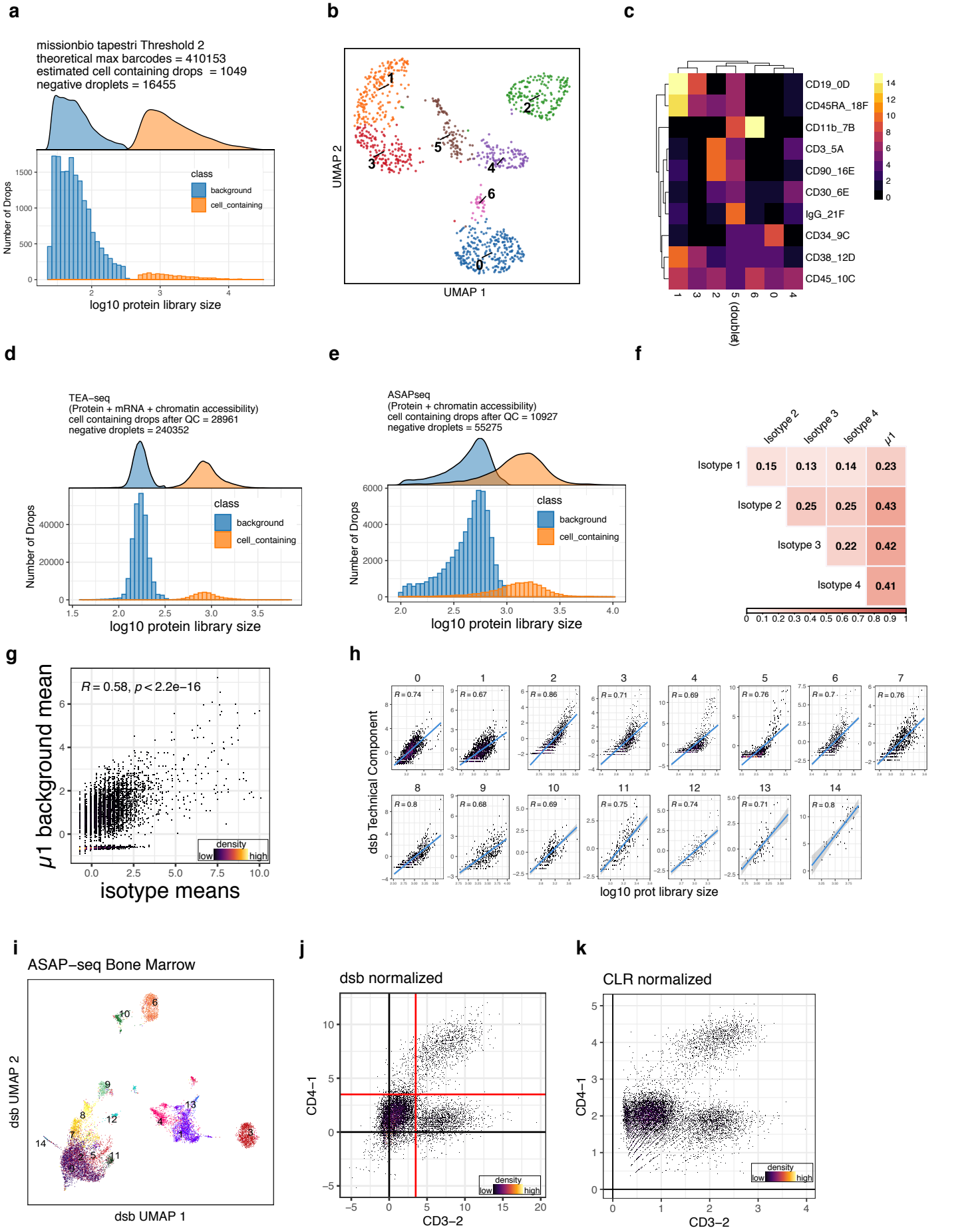
Supplementary Figure 2
Assessment of the modeling assumptions for defining each cell's background protein population mean $\mu 1$ with a k=2 component mixture model for use in the per-cell technical component regressed out of dsb normalized counts in step II–see related figures on external validation datasets (Supplementary Fig. 4). **a.** Gaussian mixture model fits (from Figs. 1d-e) partitioned by each protein-based cell cluster (clusters are the same as defined after dsb normalization in Kotliarov *et. al.* 2020). Boxplots show the median BIC with hinges at the 25th and 75th percentile and whiskers extending plus or minus 1.5 times the inter quartile range. The number of cells for each cluster: cluster 0 = 10927, 1 = 8268, 10 = 1250, 11 = 967, 12 = 853, 13 = 773, 14 = 371, 15 = 343, 16 = 292, 17 = 225, 18 = 218, 19 = 165, 2 = 6655, 20 = 137, 21 = 74, 3 = 4853, 4 = 4507, 5 = 4236, 6 = 2510, 7 = 2287, 8 = 1892, 9 = 1398. **b.** Similar to the barplot shown in Fig.1e, but partitioned by high resolution protein based cluster; cells with k = 3 as the best fit were not biased to a specific protein-based cluster. **c.** For 17% of cells with k = 3 models having the best fit (cells from Fig 1e), the difference in BIC between k = 3 vs. k = 2 and k = 2 vs. k = 1 models is shown. **d.** The distribution of Gaussian mixture model subpopulation means for k = 2 and k = 3 models for the subset of cells with k = 3 as the optimal fit (means < 15 shown to focus on μ1 distributions) shows k = 3 and k = 2 models fit similar values for μ1 in these cells. **e.** As in (d); the small minority of cells (shown in red in (b)) with k = 4 as the best fit. **f.** A single arbitrary example cell that had an optimal BIC with the k = 3 model; the distribution of inferred mixture model means is shown for the 2-subpopulation (left) and 3-subpopulation (right) model fits showing overlapping value for $\mu 1$. **g**. As in Fig.1h, using the 10X Genomics CITE-seq dataset "PBMC V3 10K" which measured only 14 surface phenotyping proteins and 3 isotype controls. The distribution of the dsb technical component as calculated using a 2 component (x-axis) vs. 3 component (y-axis) mixture model to define the μ1 parameter.


a
dsb Technical Component
log10 prot library size
count
b
positive mean μ2
background mean μ1
R = 0.62 , p < 2.2e−16
c
positive mean μ2
mean of isotype controls
R = 0.29 , p < 2.2e−16
d
100 random samples of k = 4 μ1 proteins from n = 28229 cells:
Distribution of μ1 k−sample means Pearson correlation w/ μ2 (blue)
Density
Pearson Correlation
Pearson correlation isotype controls and μ2
e
dsb Normalization
protein-specific ambient noise & cell-intrinsic technical component removed
dsb Normalization
only protein-specific ambient noise removed
CLR Transformation (across cells)
Log counts per 1e4 (library size normalization)
CD14 surface protein
CD4 surface protein
unstained cells (ground truth noise)
stained cells
f
Gap Statistic
number of clusters (k)
method
clr
dsb
g
CD14 Mono, mDC, HSC
CD8+ Memory T and NKT
CD4+ Memory T cells
B cells
NK Cells
Log2 Fold Change

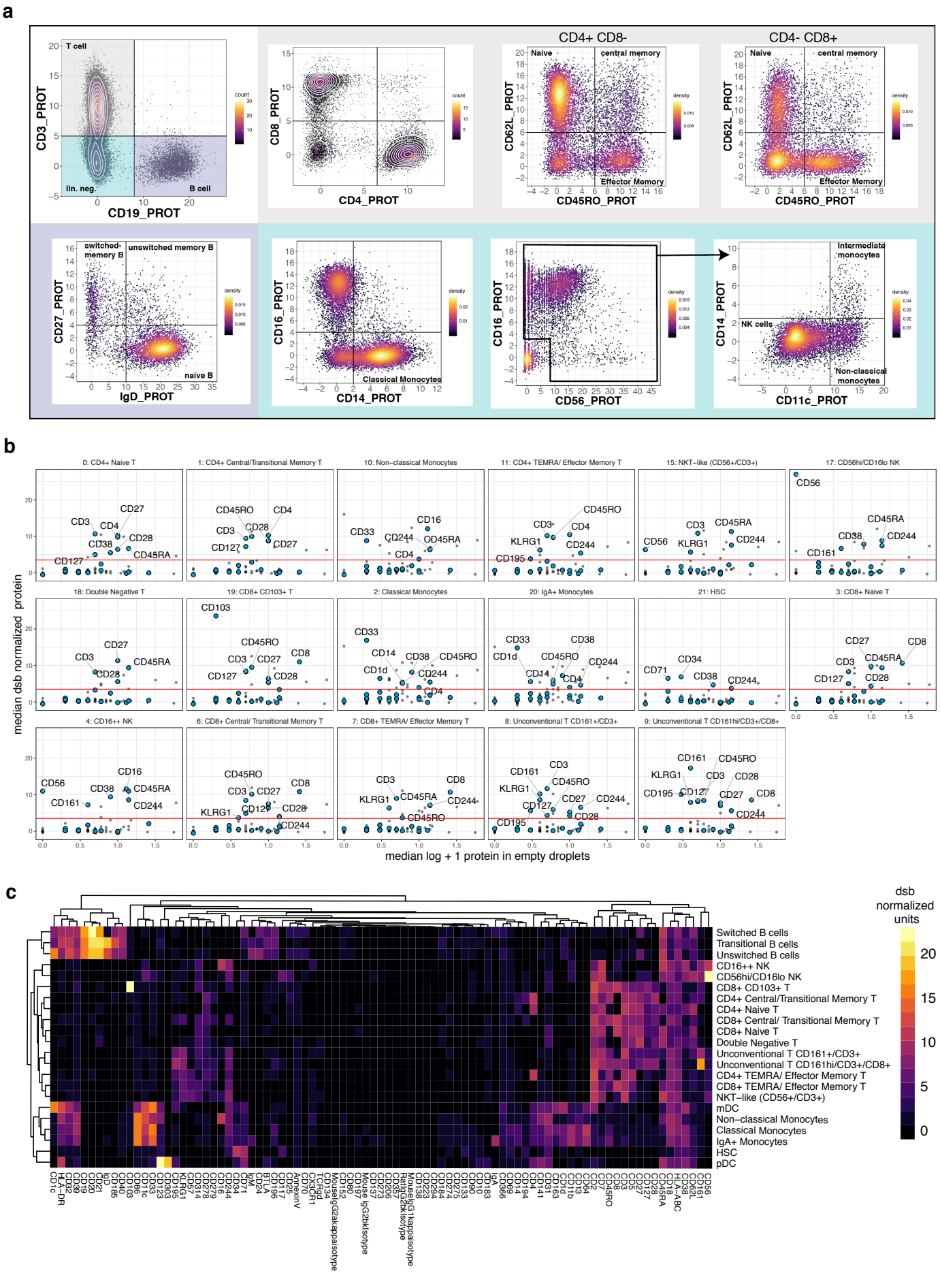
Supplementary Figure 3
a. Each cell's inferred technical component λ (y-axis) vs the cell's protein library size; panel number indicates protein-based clusters (see Methods) as shown in Supplementary Fig. 6 and Fig 3. R indicates Pearson correlation coefficient of linear fit, 95% confidence interval highlighted in grey. **b.** The Pearson correlation coefficient and p value (two sided) between μ1 and μ2 from single cell k = 2 component mixture models fit across all proteins in each cell. **c.** The average of isotype controls after dsb normalization step I (ambient correction) vs μ2 as in (b) Pearson correlation coefficient and p value (two-sided). **d**. The distribution of n=100 Pearson correlation coefficients between each cell's μ2 and 100 random samples of k=4 μ1 proteins from each single cell (blue) shaded region is the 50% highest density interval, red line is the Pearson correlation coefficient of μ2 and the mean of isotype controls in each cell from batch 1 (28,229 cells). **e.** Single cell protein expression of CD4 vs. CD14 normalized by different methods. Contour lines in red are the distribution of CD4 and CD14 in unstained control cells after normalization in the exact same way as the stained cells in black within each panel, including dsb normalization using the same empty droplets for ambient correction of the unstained cells. Outlier cells (less than 0.3% of total cells in any panel) are removed to focus on the three main cell populations. The default implementation of dsb using steps I and II (top left panel) and CLR across cells are shaded in blue and grey respectively as these methods are further compared in subsequent panels and in Figs. 4 and 5. **f.** The Gap Statistic (see Methods) for different number of clusters (k) obtained using the k-medoids clustering algorithm on normalized protein values from dsb vs. CLR (across cells), bars are standard errors of the gap statistic calculated by the clusGap R function. **g.** Log fold-change estimates from differential expression analysis of proteins for each major cell type shown in comparison with the rest of the cell types (blue – dsb, grey – CLR across cells).


a
10X PBMC 10K V3 Threshold 2
theoretical max barcodes = 6794880
cell containing drops after QC = 7087
negative droplets = 69481
Number of Drops
class
background
cell_containing
log10 protein library size
b
Isotype 2
Isotype 3
μ1
Isotype 1
0.43
0.31
0.55
Isotype 2
0.35
0.58
Isotype 3
0.47
c
R = 0.77 , p < 2.2e−16
μ1 background mean
isotype means
d
dsb Technical Component
log10 prot library size
e
UMAP 2
UMAP 1
f
CD45RA_TotalSeqB
CD4_TotalSeqB
CD3_TotalSeqB
CD127_TotalSeqB
CD19_TotalSeqB
CD14_TotalSeqB
CD45RO_TotalSeqB
CD8a_TotalSeqB
CD56_TotalSeqB
CD16_TotalSeqB
TIGIT_TotalSeqB
CD25_TotalSeqB
PD−1_TotalSeqB
IgG2a_control_TotalSeqB
IgG1_control_TotalSeqB
CD15_TotalSeqB
IgG2b_control_TotalSeqB
g
CD14
CD4
density
low
high
h
10X PBMC5k V3
theoretical max barcodes = 6794880
cell containing drops after QC = 3481
negative droplets after QC = 67204
Number of Drops
class
background
cell_containing
log10 protein library size
i
Isotype 2
Isotype 3
μ1
Isotype 1
0.22
0.18
0.42
Isotype 2
0.21
0.48
Isotype 3
0.46
j
R = 0.74 , p < 2.2e−16
μ1 background mean
isotype means
k
dsb Technical Component
log10 prot library size
l
UMAP 2
UMAP 1
m
CD19_TotalSeqB
CD20_TotalSeqB
CD16_TotalSeqB
CD56_TotalSeqB
CD335_TotalSeqB
TIGIT_TotalSeqB
PD−1_TotalSeqB
CD25_TotalSeqB
CD15_TotalSeqB
CD197_TotalSeqB
CD69_TotalSeqB
IgG2b_control_TotalSeqB
CD274_TotalSeqB
IgG1_control_TotalSeqB
CD80_TotalSeqB
CD137_TotalSeqB
IgG2a_control_TotalSeqB
HLA−DR_TotalSeqB
CD11b_TotalSeqB
CD14_TotalSeqB
CD86_TotalSeqB
CD45RO_TotalSeqB
CD3_TotalSeqB
CD278_TotalSeqB
CD28_TotalSeqB
CD127_TotalSeqB
CD4_TotalSeqB
CD62L_TotalSeqB
CD8a_TotalSeqB
CD27_TotalSeqB
CD45RA_TotalSeqB
n
CD14
CD4
density
low
high
o
10X PBMC 5' 5K V2
theoretical max barcodes = 737280
cell containing drops after QC = 7089
negative droplets after QC = 55541
Number of Drops
class
background
cell_containing
log10 protein library size
p
Isotype 2
Isotype 3
μ1
Isotype 1
0.49
0.34
0.66
Isotype 2
0.34
0.64
Isotype 3
0.59
q
R = 0.86 , p < 2.2e−16
μ1 background mean
isotype means
r
dsb Technical Component
log10 prot library size
s
UMAP 2
UMAP 1
t
CD8a_TotalSeqC
CD3_TotalSeqC
CD127_TotalSeqC
CD19_TotalSeqC
CD14_TotalSeqC
CD45RO_TotalSeqC
CD25_TotalSeqC
TIGIT_TotalSeqC
IgG2b_control_TotalSeqC
IgG1_control_TotalSeqC
IgG2a_control_TotalSeqC
PD−1_TotalSeqC
CD15_TotalSeqC
CD16_TotalSeqC
CD56_TotalSeqC
CD45RA_TotalSeqC
CD4_TotalSeqC
u
CD14
CD4
density
low
high

Supplementary Figure 4
Panels as shown in Fig. 2 (10X Genomics dataset “PBMC 5k” Next Gem assay) for additional 10X genomics datasets using different assays and protein panels. **a-g** “PBMC 10k” V3 assay, **h-n** “PBMC 5k” V3 assay and **o-u** “PBMC 5k” 5 prime V2 assay. 95% confidence intervals of linear model fits (d,k,r) in grey. Pearson correlation coefficients and p values (two sided) are shown (c, j, q).


a
missionbio tapestri Threshold 2
theoretical max barcodes = 410153
estimated cell containing drops = 1049
negative droplets = 16455
class
background
cell_containing
Number of Drops
log10 protein library size
b
UMAP 2
UMAP 1
c
CD19_0D
CD45RA_18F
CD11b_7B
CD3_5A
CD90_16E
CD30_6E
IgG_21F
CD34_9C
CD38_12D
CD45_10C
5 (doublet)
d
TEA–seq
(Protein + mRNA + chromatin accessibility)
cell containing drops after QC = 28961
negative droplets = 240352
e
ASAPseq
(Protein + chromatin accessibility)
cell containing drops after QC = 10927
negative droplets = 55275
f
Isotype 2
Isotype 3
Isotype 4
μ1
Isotype 1
0.15
0.13
0.14
0.23
0.25
0.25
0.43
0.22
0.42
0.41
g
R = 0.58, p < 2.2e–16
μ1 background mean
isotype means
density
low
high
h
dsb Technical Component
log10 prot library size
i
ASAP–seq Bone Marrow
dsb UMAP 2
dsb UMAP 1
j
dsb normalized
CD4–1
CD3–2
k
CLR normalized
CD4–1
CD3–2

Supplementary Figure 5
a. A mixture of n=4 leukemia cell lines from example data generated via the Mission Bio 'Tapestri" platform for simultaneous "proteogenomic" assessment of surface proteins and DNA**.** The protein library size (total UMI) distribution was used to distinguish between cell-containing and empty droplets without cells. **b.** UMAP analysis based on dsb normalized values; cells are labeled by graph-based cluster identity. **c.** heatmap of the average expression of each dsb-normalized protein in each cluster. The range of values is on the same scale for all proteins, ranging from less than 0 to 14, corresponding to 14 standard deviations from the average background level estimated using empty droplets–cell-to-cell technical variations were not inferred by calculating the technical component for each cell (step II of dsb) in this dataset due to the small number of proteins profiled (n=10, see Supplementary Note and Methods). **d.** As in (a) for TEA-seq and **e.** ASAP-seq datasets. Cell-containing droplets defined by the QC pipeline from Swanson *et. al.* and Mimitau *et. al.*, respectively; note that only protein was used to estimate background from the subset of droplets that did not meet cell QC for the ASAP-seq dataset (see methods). **f.** As in Fig. 1f, correlation matrix of variables comprising the dsb technical component. **g.** As in Fig.1g, isotype control mean vs. background mean per cell. Pearson correlation coefficient and p value (two sided) is shown. **h**. As in Supplementary Fig. 3a, relationship between protein library size and the dsb technical component. Linear trend shown in blue with 95% confidence intervals in grey. **i.** UMAP projection and clusters based on dsb normalized protein values. **j.** Biaxial plot of CD3 vs. CD4 with the dsb threshold of 3.5 shown. **k.** As in (j) but with data normalized using the CLR transformation (across cells).

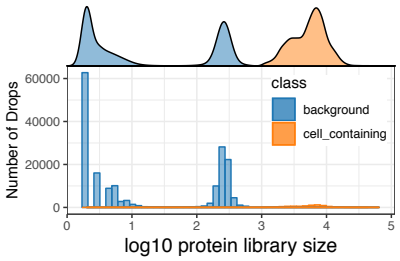

a
T cell
lin. neg.
B cell
CD3_PROT
CD19_PROT
CD4+ CD8-
CD8_PROT
CD4_PROT
CD4- CD8+
Naive
central memory
Effector Memory
CD62L_PROT
CD45RO_PROT
switched-memory B
unswitched memory B
naive B
CD27_PROT
IgD_PROT
Classical Monocytes
CD16_PROT
CD14_PROT
CD56_PROT
Intermediate monocytes
NK cells
Non-classical monocytes
CD11c_PROT
b
median dsb normalized protein
median log + 1 protein in empty droplets
0: CD4+ Naive T
1: CD4+ Central/Transitional Memory T
10: Non−classical Monocytes
11: CD4+ TEMRA/ Effector Memory T
15: NKT−like (CD56+/CD3+)
17: CD56hi/CD16lo NK
18: Double Negative T
19: CD8+ CD103+ T
2: Classical Monocytes
20: IgA+ Monocytes
21: HSC
3: CD8+ Naive T
4: CD16++ NK
6: CD8+ Central/ Transitional Memory T
7: CD8+ TEMRA/ Effector Memory T
8: Unconventional T CD161+/CD3+
9: Unconventional T CD161hi/CD3+/CD8+
c
dsb normalized units
Switched B cells
Transitional B cells
Unswitched B cells
CD16++ NK
CD56hi/CD16lo NK
CD8+ CD103+ T
CD4+ Central/Transitional Memory T
CD4+ Naive T
CD8+ Central/ Transitional Memory T
CD8+ Naive T
Double Negative T
Unconventional T CD161+/CD3+
Unconventional T CD161hi/CD3+/CD8+
CD4+ TEMRA/ Effector Memory T
CD8+ TEMRA/ Effector Memory T
NKT−like (CD56+/CD3+)
mDC
Non−classical Monocytes
Classical Monocytes
IgA+ Monocytes
HSC
pDC

Supplementary Figure 6
a. Biaxial gating strategy for identifying major immune cell subsets with dsb normalized values. Grey = T cells, Blue = Monocytes, Purple = B cells. **b.** As in Fig. 3e, the average log transformed protein count in empty droplets (x-axis) vs the average dsb normalized values (y-axis) for each protein-based cell cluster–the threshold above which proteins are annotated in the plot is 3.5 corresponding to 3.5 standard deviations above expected noise +/- the technical component correction applied in step II (see methods). In each plot the same subset of proteins is highlighted in blue for comparison of individual marker values between clusters; proteins highlighted in blue are CD1d, CD1c, CD14, CD103, CD16, CD3, CD4, CD8, CD28, CD161, CD45RO, CD45RA, CD33, CD56, CD71, CD27, CD244, KLRG1, CD195, CD38, CD127, CD16, CD34. When the protein value is above the 3.5 threshold, it is labeled with the protein name in each individual panel. **c.** Heatmap of average dsb protein normalized expression in each cluster.

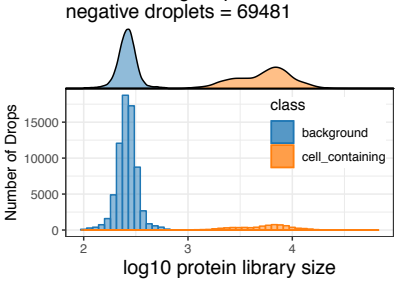
a

define background droplets:

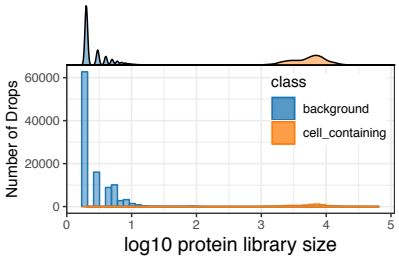
10X PBMC 10K V3 Threshold 1
theoretical max barcodes = 6794880
cell containing drops after QC = 7087
negative droplets = 176874



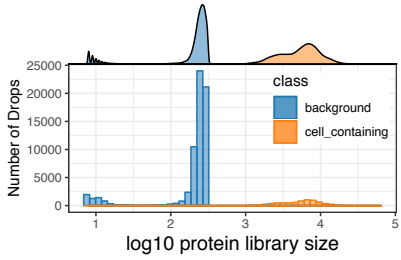
10X PBMC 10K V3 Threshold 2
theoretical max barcodes = 6794880
cell containing drops after QC = 7087
negative droplets = 69481



10X PBMC 10K V3 Threshold 3
theoretical max barcodes = 6794880
cell containing drops after QC = 7087
negative droplets = 107382

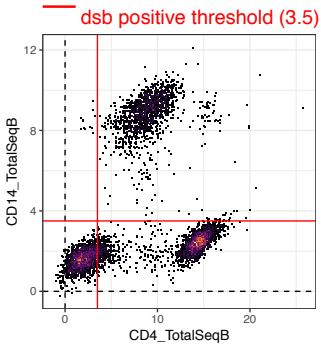


10X PBMC 10K V3 Threshold 4
theoretical max barcodes = 6794880
cell containing drops after QC = 7087
negative droplets = 66157

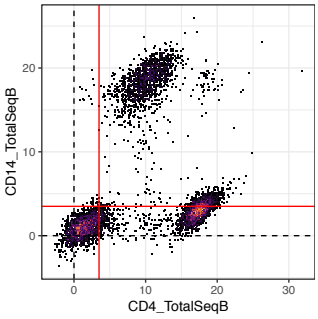


b

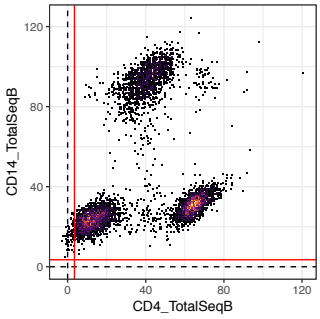
Normalize with dsb using different subsets of background
for protein-specific ambient correction (dsb step I)



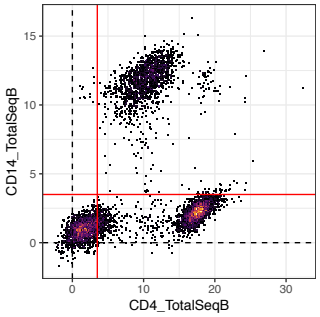
background droplets with > 1
protein UMI used for dsb normalization
0 > log10(protUMI) < 2.8



simulation of major background
population used by dsb only
2 < log10(protUMI) < 2.8

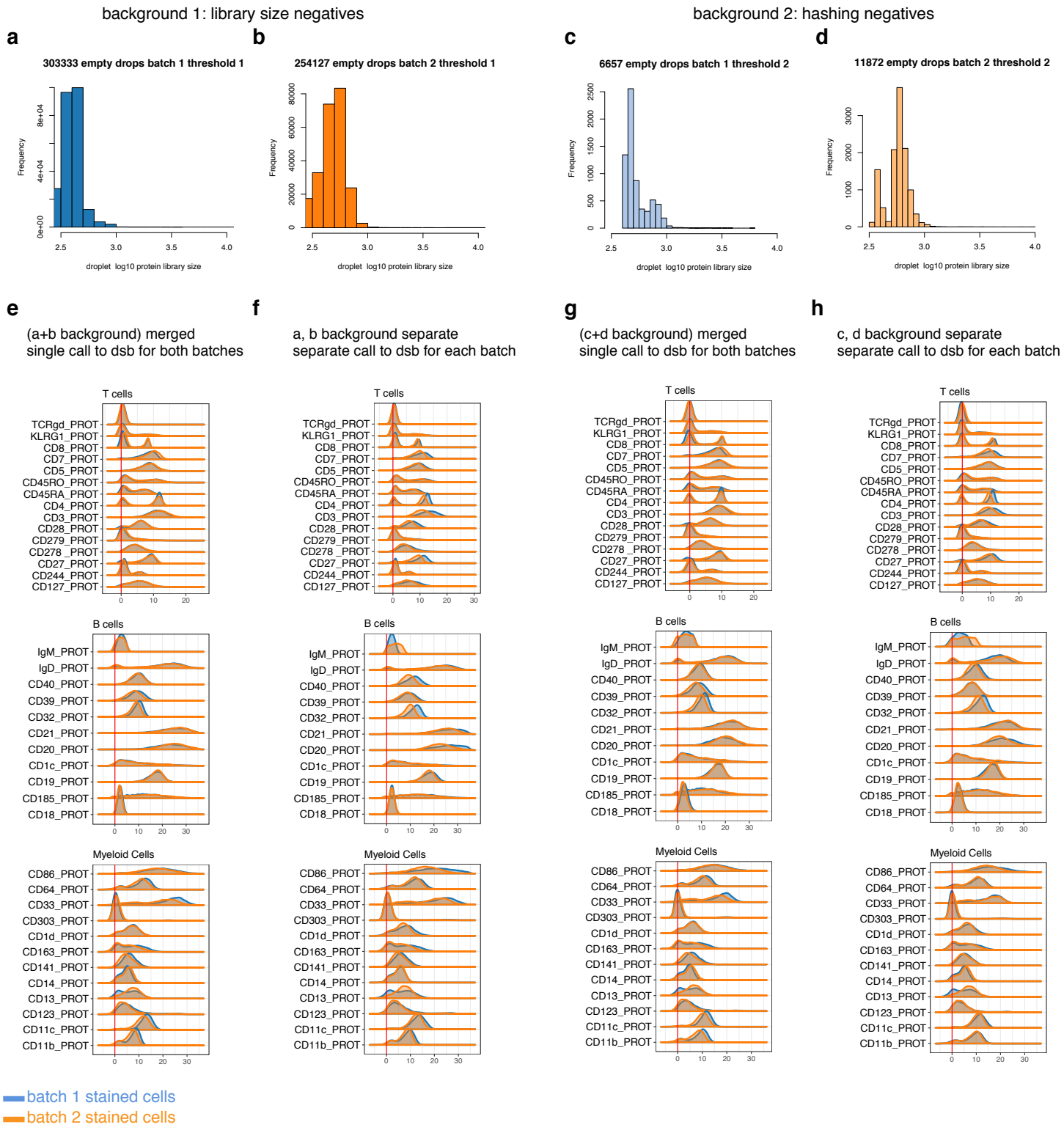


simulation of major background
droplet population missing
0 < log10(protUMI) < 2

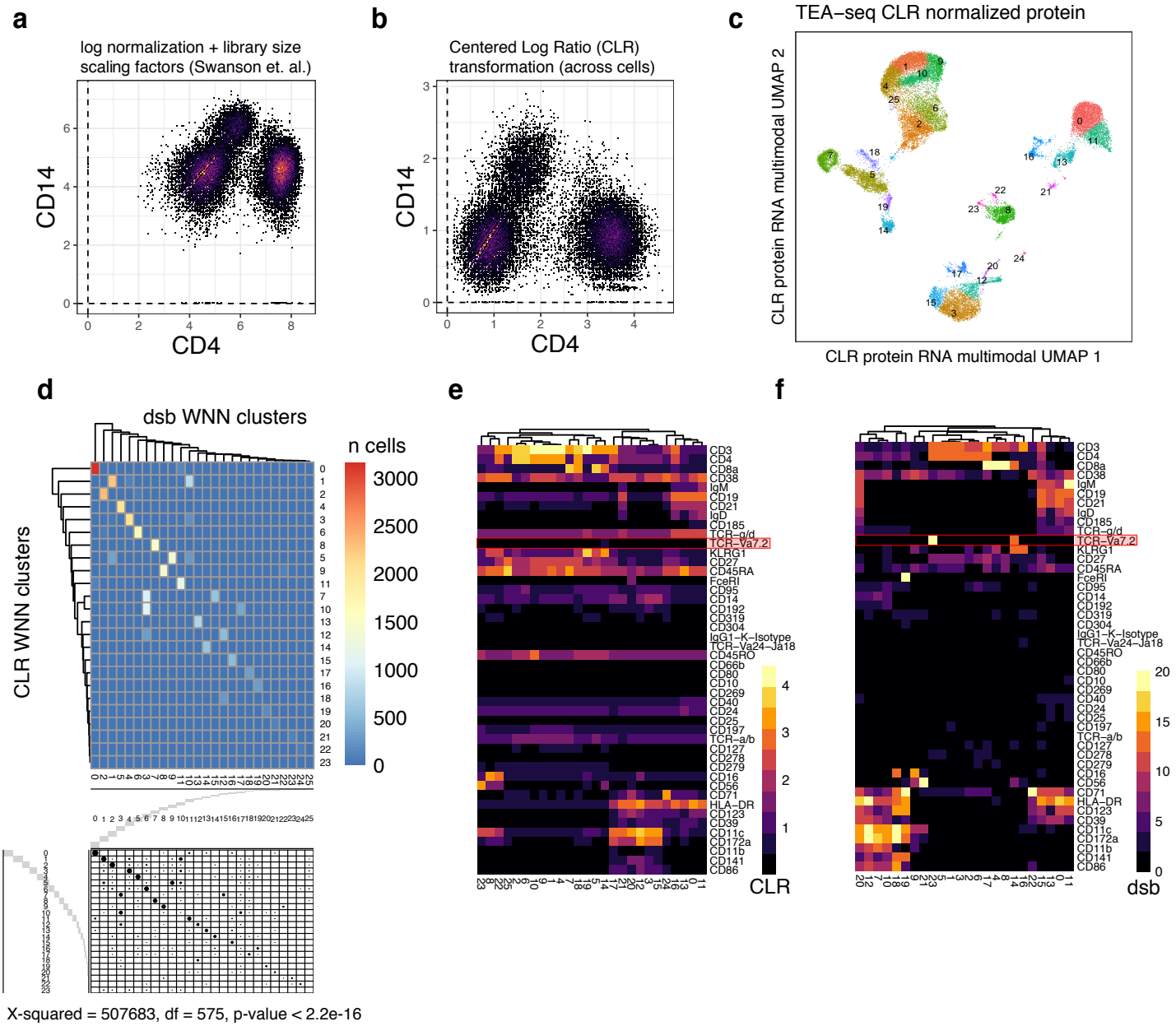


simulation of removal of the major low
population and highest (90th percentile)
of background droplets
0.9 < log10(protUMI) < 2.5

Supplementary Figure 7
Robustness of dsb normalized values to different definitions of background droplets. **a.** Distribution of protein library size for the 10X genomics Chromium Version 3 "PBMC 10K" dataset which had a bimodal distribution for the non-cell-containing droplets shown in blue. In each row, a different threshold based on the protein library size was used to define background droplets, which were then used to normalize the same population of cell-containing droplets (shown as the orange distribution) with the dsb package. **b.** The dsb normalized values are shown for canonical protein-based phenotypes with biaxial scatterplots. The scale of the 3rd row is negatively impacted by eliminating the major empty droplet background peak with greater mean value and only using the empty droplet background peak with very low mean protein library size.


background 1: library size negatives
background 2: hashing negatives
a
303333 empty drops batch 1 threshold 1
b
254127 empty drops batch 2 threshold 1
c
6657 empty drops batch 1 threshold 2
d
11872 empty drops batch 2 threshold 2
Frequency
droplet log10 protein library size
e
(a+b background) merged
single call to dsb for both batches
f
a, b background separate
separate call to dsb for each batch
g
(c+d background) merged
single call to dsb for both batches
h
c, d background separate
separate call to dsb for each batch
T cells
TCRgd_PROT
KLRG1_PROT
CD8_PROT
CD7_PROT
CD5_PROT
CD45RO_PROT
CD45RA_PROT
CD4_PROT
CD3_PROT
CD28_PROT
CD279_PROT
CD278_PROT
CD27_PROT
CD244_PROT
CD127_PROT
B cells
IgM_PROT
IgD_PROT
CD40_PROT
CD39_PROT
CD32_PROT
CD21_PROT
CD20_PROT
CD1c_PROT
CD19_PROT
CD185_PROT
CD18_PROT
Myeloid Cells
CD86_PROT
CD64_PROT
CD33_PROT
CD303_PROT
CD1d_PROT
CD163_PROT
CD141_PROT
CD14_PROT
CD13_PROT
CD123_PROT
CD11c_PROT
CD11b_PROT
batch 1 stained cells
batch 2 stained cells

Supplementary Figure 8
Stability of dsb normalized values when processing multiple batches in a single normalization vs normalizing each batch separately, both using two definitions of background droplets with the dsb package. **a–d** show protein library size distributions of background droplets defined using either the protein library size distribution alone or droplets defined as negative during demultiplexing (see Supplemental note) across n = 2 batches. The raw Cell Ranger outputs from each staining batch of cells were split across n=6 lanes per batch of the 10X Chromium instrument and for each definition of background, the dsb results for merged vs split batch normalization are shown in **e-h**.



X-squared = 507683, df = 575, p-value < 2.2e-16

Supplementary Figure 9
Analysis of TEA-seq (transcriptome, epitopes and accessibility) tri-modal single cell assay data. **a.** TEA-seq data normalized by Library size based normalization (as in Swanson et. al.), and **b.** CLR across cells. **c.** UMAP plot of single cells and clusters derived by WNN joint mRNA-protein clustering with data normalized using CLR (see Fig. 4b for dsb normalized data). **d.** Contingency of clustering results between joint mRNA and protein Weighted Nearest Neighbor (WNN) clustering with CLR normalized (rows) or dsb normalized (columns) values as input to the protein matrix. The bottom margin shows the same data as circles with area proportional to frequency to show clusters with cell assignment differences. The average protein expression profiles of the clusters from **e.** CLR and **f.** dsb for protein normalization are shown as heatmaps.