Published online 9 September 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 17* ***9633–9647***
https://doi.org/10.1093/nar/gkab745

SPaRTAN, a computational framework for linking cell-surface receptors to transcriptional regulators

Xiaojun Ma[1,6,†], Ashwin Somasundaram[2,6,†], Zengbiao Qi[6], Douglas J. Hartman[3,6], Harinder Singh[4] and Hatice Ulku Osmanbeyoglu[1,5,6,*]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15206, USA, [2]Department of Medicine, Division of Hematology/Oncology, University of Pittsburgh, Pittsburgh, PA 15213, USA, [3]Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA, [4]Center for Systems Immunology and Department of Immunology, University of Pittsburgh, Pittsburgh, PA 15213, USA, [5]Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA 15261, USA and [6]UPMC Hillman Cancer Center, Pittsburgh, PA 15213, USA

Received March 17, 2021; Revised August 09, 2021; Editorial Decision August 12, 2021; Accepted September 06, 2021

ABSTRACT

The identity and functions of specialized cell types are dependent on the complex interplay between signaling and transcriptional networks. Recently single-cell technologies have been developed that enable simultaneous quantitative analysis of cell-surface receptor expression with transcriptional states. To date, these datasets have not been used to systematically develop cell-context-specific maps of the interface between signaling and transcriptional regulators orchestrating cellular identity and function. We present SPaRTAN (Single-cell Proteomic and RNA based Transcription factor Activity Network), a computational method to link cell-surface receptors to transcription factors (TFs) by exploiting cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) datasets with cis-regulatory information. SPaRTAN is applied to immune cell types in the blood to predict the coupling of signaling receptors with cell context-specific TFs. Selected predictions are validated by prior knowledge and flow cytometry analyses. SPaRTAN is then used to predict the signaling coupled TF states of tumor infiltrating CD8$^+$ T cells in malignant peritoneal and pleural mesotheliomas. SPaRTAN enhances the utility of CITE-seq datasets to uncover TF and cell-surface receptor relationships in diverse cellular states.

INTRODUCTION

The reciprocal interplay between complex signaling inputs and transcriptional responses dictate the generation of distinct cell types and their specialized functions. Dysregulation of this interplay leads to the development and progression of disease, most clearly delineated in the context of certain cancers, chronic infections and autoimmune diseases. Understanding these dynamic programs at the single-cell level represents a formidable challenge. Emerging single-cell genomic technologies (1–4) provide a transformative platform to characterize, in a comprehensive and unbiased manner, the full range of cell types and their genomic programming in health and disease.

The computational prediction of gene regulatory programs based on single-cell genomic datasets is a relatively new field. There is still a large methodological gap between generating single-cell datasets and delineating cell-specific regulatory programs orchestrating cellular identity and function. Early gene regulatory program inference methods use single-cell RNA-seq (scRNA-seq) data alone or in combination with TF motifs in annotated promoter regions (5–9). These methods primarily depend on co-expression of TFs and their potential target genes (6) and thus are not suitable for many TFs whose transcripts are expressed at low levels or whose activities are post-transcriptionally regulated. Moreover, co-expression may not always imply co-regulation. With the recent availability of single-cell epigenomic datasets, the tools of regulatory genomics are being applied to infer TFs associated with accessible chromatin regions (10) at both promoter-proximal as well as distal regions and in turn with gene expression (11). However, these approaches do not comprehensively consider the relationships between signaling systems (e.g. from proteomic data) and transcriptional states of individual cells. Recent breakthroughs in single-cell genomics have linked single-cell gene expression data with quantitative protein measurements using index sorting (12) and barcoded antibodies (1,2), in particular *cellular indexing of transcriptomes and epitopes by sequencing* (CITE-seq) (1). CITE-seq adds a step in which barcoded antibodies—a second set of barcodes—are incubated with the single-cell

[*]To whom correspondence should be addressed. Tel: +1 412 623 7789; Email: osmanbeyogluhu@pitt.edu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

suspension into droplet based scRNA-seq protocol. These barcodes, which have polyA tails, are then linked to the barcodes from beads at the same time mRNAs are linked. Reads containing barcodes associated with each bead are separated by cell. Then, reads that align to transcripts are used to quantify mRNA levels while those from barcoded antibodies (antibody-derived tags (ADTs)) are used to quantify protein levels. To date, the datasets generated by this powerful platform have not been used to link the expression of cell surface proteins, for example, signaling receptors, with the activities of TFs and gene expression programs in individual cells.

Here, we describe a computational framework for exploiting single-cell proteomic (scADT-seq) and corresponding single-cell transcriptomic (scRNA-seq) datasets, both obtained using CITE-seq, to link expression of surface proteins with inferred TF activities. Our framework, SPaRTAN (**S**ingle-cell **P**roteomic **a**nd **R**NA based **Tr**anscription factor **A**ctivity **N**etwork), advances our prior algorithmic approach based on bulk tumor datasets (13–15). SPaRTAN model views expression of surface proteins (ADT counts) as a proxy of their activities; signaling emanating from these proteins converges on particular TFs, whose activities, in turn, regulate the expression of their target genes (Figure 1). More specifically, we use a regularized bilinear regression algorithm called affinity regression (AR) (16) to learn a cell-type specific interaction matrix between upstream cell-surface receptor proteins and downstream TFs that predicts target gene expression. The trained SPaRTAN model can then infer the TF activity given a cell's surface protein expression profile or infer the cell-surface receptor expression given a cell's gene expression profile. We apply and experimentally test SPaRTAN using CITE-seq datasets from peripheral blood mononuclear cells (PBMCs) and then illustrate its broader utility by predicting signaling coupled TF activities in tumor infiltrating CD8$^+$ T cells in the context of malignant peritoneal and pleural mesothelioma.

MATERIALS AND METHODS

Data and preprocessing

CITE-seq data for 5k (Chemistry v3), 5k (Nextgen) PBMC obtained from 10x Genomics website (Supplementary Table S1). A total of ∼5000 cells from a healthy donor were stained with 29 TotalSeq-B antibodies, including CD3, CD4, CD8a, CD11b, CD14, CD15, CD16, CD19, CD20, CD25, CD27, CD28, CD34, CD45RA, CD45RO, CD56, CD62L, CD69, CD80, CD86, CD127, CD137, CD197, CD274, CD278, CD335, PD-1, HLA-DR and TIGIT. Cell-matched scRNA-seq data are available. To further evaluate the validity of our method, we also generated an in-house CITE-seq dataset of human malignant peritoneal and pleural mesothelioma under IRB approval from the University of Pittsburgh. Cells were stained with TotalSeq-C from BioLegend and are prepared using the 10x Genomics platform with Gel Bead Kit V2 (as described below). Forty-six surface markers are measured for every cell: adt-CD274 (B7-H1, PD-L1), adt-CD273, adt-CD30, adt-CD40, adt-CD56, adt-CD19, adt-CD14, adt-CD11c, adt-CD117, adt-CD123, adt-CD194 (CCR4), adt-CD4, adt-CD25, adt-CD279, adt-TIGIT, adt-CD20, adt-CD195 (CCR5), adt-CD185 (CXCR5), adt-CD103 (Integrin αE), adt-CD69, adt-CD62L, adt-CD197, adt-CD161, adt-CD152 (CTLA-4), adt-CD223 (LAG-3), adt-CD27, adt-CD95, adt-CD134 (OX40), adt-HLA-DR, adt-CD1c, adt-CD11b, adt-CD141, adt-CD314, adt-CD66b, adt-CD366, adt-CD278, adt-CD39, adt-KLRG1, adt-CD137, adt-CD254, adt-CD357, adt-CD28, adt-CD38, adt-CD127 (IL-7Rα), adt-CD15 and adt-TCRVdelta2.

Normalization and initial explanatory analysis of CITE-seq datasets were performed using the Seurat R package version 3.1.5 (17). During quality control, we excluded cells with <300 and >5000 expressed genes, the latter to avoid doublets. Antibody-derived tags (ADTs) for each cell were normalized using a centered log ratio (CLR) transformation across cells. We performed log-normalization for all scRNA-seq datasets using a size factor of 10,000 molecules for each cell. Seurat 'FindClusters' was applied to the first 50 principal components, with the resolution parameter set to 1. Cell labels were assigned using marker genes' protein and gene expression levels.

To construct the TF–target gene prior matrix, we downloaded a gene set resource containing TF target–gene interactions from DoRothEA (18). Those interactions were curated and collected from different types of evidence such as literature curated resources, ChIP-seq peaks, TF binding site motifs, and interactions inferred directly from gene expression. This TF–target gene prior matrix (**D**) defines a candidate set of associations between TFs and target genes. Further, we filtered TFs that were not expressed across all cell-types. Processed data files have also been made available at the supplementary website for the paper (see URLs).

Tissue processing of malignant mesothelioma

Tumor samples were washed in RPMI containing antibiotics such as amphotericin B and penicillin–streptomycin for 30 min followed by mechanical and enzymatic digestion and further passage via a 100 μm filter. Isolated tumor infiltrating lymphocytes (TIL) were then washed in RPMI media twice, and stained with aforementioned TotalSeqC antibodies, and CD45-PE, EpCAM, and cell viability dyes. After washing, immune cells were sorted based on CD45 and utilized for sequencing library preparation.

Simultaneous protein and transcriptomic single cell profiling of malignant mesotheliomas via CITE-seq

Combined surface protein and mRNA expression single cell analysis was performed using CITE-seq methodology as previously described (1).

Generation of scRNAseq libraries: Live CD45$^+$, EpCAM$^-$ cells (i.e. all immune cells) and live EpCAM$^+$ (tumor cells) were sorted from tumor tissue. Single-cell libraries were generated utilizing the chromium single-cell 5′ Reagent (V2 chemistry). Briefly, sorted cells were re-suspended in PBS (0.04% BSA; Sigma) and then loaded into the 10× Controller for droplet generation, targeting recovery of 5000 cells per sample. Cells were then lysed and reverse transcription was performed within the droplets and cDNA was isolated and amplified in bulk with 12 cycles of PCR. Amplified libraries were then size selected

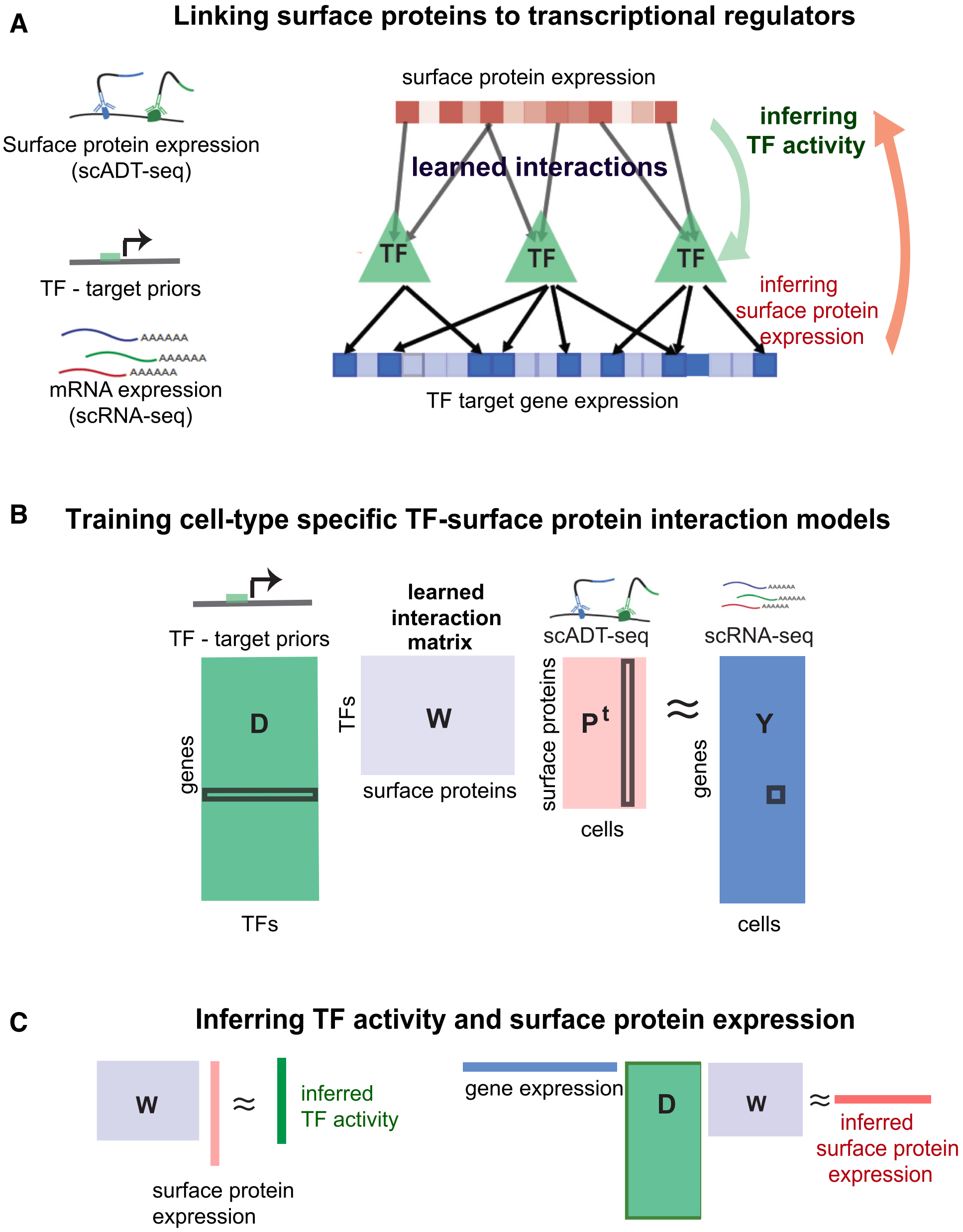


Figure 1. Integrative computational model linking cell-surface receptors to transcriptional regulators—application to peripheral blood mononuclear cells (PBMC). (**A**) Our integrative model (SPaRTAN, Single-cell Proteomic and RNA based Transcription factor Activity Network) utilizes single-cell multi-omics data from cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) datasets and infers the flow of information from cell-surface receptors to transcription factors (TFs) to target genes by learning interactions between cell surface receptors and TFs that best predict target gene expression. (**B**) SPaRTAN trains on mRNA and surface protein expression data from a set of cells, along with curated TF target-gene interactions, to learn a model that links upstream signaling to downstream transcriptional response. Specifically, the algorithm learns a weight matrix **W** between cell-surface proteins and TFs that predicts gene expression (**Y**) from cell-specific surface protein levels (**P**) and TF target-gene interactions (**D**) by solving the bilinear regression problem shown. (**C**) Using the trained interaction matrix (**W**), one can predict TF activities from the cell surface protein expression profile, or protein activities from the cellular mRNA expression data and the TF-target gene hit matrix.

utilizing SPRIselect beads, and adapters were ligated followed by sample indices. After another round of SPRIselect purification, a KAPA DNA Quantification PCR determined the concentration of libraries. The supernatant after first SPRIselect beads, containing ADTs, was used to generate an ADT library.

Sequencing of single-cell libraries: Libraries were diluted to 2nM and pooled for sequencing by NextSeq500/550 high-output v2 kits (UPMC Genomic Center) for 150 cycles (parameters: Read 1: 26 cycles; i7 index: 8 cycles, Read 2: 98 cycles). The prepared assay is subsequently sequenced on a NextSeq500/550 with a depth of 50K reads per cell. Raw sequence data were processed via CellRanger 3.0 (10× Genomics) and aligned to GRCh38 to generate UMI matrix for the downstream analysis. Cell barcodes with fewer than 3 UMI counts in 1% of cells were removed.

Training cell-type specific SPaRTAN models

We trained cell-type specific SPaRTAN models using an affinity regression (AR) algorithm for efficiently solving a regularized bilinear regression problem (13,16), defined here as follows. For a data set of M cells profiled using scRNA-seq across N genes, we let $\mathbf{Y} \in \mathrm{R}^{\mathrm{NxM}}$ be the log-normalized gene expression matrix where each column of **Y** corresponds to a cell. We defined each gene's TF attributes in a matrix $\mathbf{D} \in \mathrm{R}^{\mathrm{NxQ}}$, where each row represents a gene and each column is a binary vector representing the target genes of a TF. We used curated TF target-gene interactions (**D**) (18) to determine the set of TFs that potentially regulate each gene. We defined the cell surface protein attributes of cells as a matrix $\mathbf{P} \in \mathrm{R}^{\mathrm{MxS}}$ where each row represents a cell and each column represents log-normalized surface protein expression profile for the cell based on scADT-seq. To improve computational stability, columns of **P**, **Y** and **D** were all normalized to have a unit norm before training. Next, we set up a bilinear regression problem to learn the weight (interaction) matrix $\mathbf{W} \in \mathrm{R}^{\mathrm{QxS}}$ between TFs and cell surface proteins that predicts TF-target gene expression by

$$\mathbf{DWP}^{\mathrm{T}} \sim \mathbf{Y}. \tag{1}$$

An illustration of the model is given in Figure 1B. To find the optimal solution of the Equation (1) that has the minimal residual, one can solve the following optimization problem:

$$\mathrm{argmin}_{\mathrm{W}} \|\mathbf{DWP}^{\mathrm{T}} - \mathbf{Y}\|_2^2 \tag{2}$$

where **D**, **P** and **Y** are known. For the ease of computation, we transformed the system to an equivalent linear system of equations by reformulating the matrix products as Kronecker products:

$$\mathbf{DWP}^{\mathrm{T}} \approx \mathbf{Y} \Leftrightarrow (\mathbf{P} \otimes \mathbf{D})\mathrm{vec}(\mathbf{W}) \approx \mathrm{vec}(\mathbf{Y}) \tag{3}$$

where ⊗ is a Kronecker product, and vec(.) is a vectorizing operator that stacks a matrix and produces a vector, yielding a standard (large-scale) regression problem. Since the number of samples (cells) and features (genes) is large, we reduced Equation (3) to a smaller system of equations by left-multiplication of by $\mathbf{Y}^{\mathrm{T}}$:

$$\mathbf{Y}^{\mathrm{T}}\mathbf{DWP}^{\mathrm{T}} \approx \mathbf{Y}^{\mathrm{T}}\mathbf{Y} \Leftrightarrow (\mathbf{P} \otimes \mathbf{Y}^{\mathrm{T}}\mathbf{D})\mathrm{vec}(\mathbf{W}) \approx \mathrm{vec}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) \tag{4}$$

and the corresponding unregularized optimization problem has form below:

$$\mathrm{argmin}_{\mathrm{W}}\left(\|\mathrm{vec}\left(\mathbf{Y}^{T}\mathbf{Y}\right) - \mathbf{P} \otimes \mathbf{Y}^{T}\mathbf{D}\right)\mathrm{vec}(\mathbf{W})\|_2^2 \tag{5}$$

The multiplication in Equation (4) creates a new output space as products $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, which effectively measures the distance between the gene expression profiles of all pairs of cells. Pelossof *et al.* (16) has shown this compressed affinity regression reduces a large-scale problem to a very compact approximate problem and effectively learns a model **W** that predicts the similarity between any pairs of samples.

To avoid overfitting, we added additional regularizes to Equation (5):

$$\mathrm{argmin}_{\mathrm{W}}\left(\|\mathrm{vec}\left(\mathbf{Y}^{T}\mathbf{Y}\right) - \mathbf{P} \otimes \mathbf{Y}^{T}\mathbf{D})\mathrm{vec}(\mathbf{W})\|_2^2 + \lambda_2\|\mathbf{W}\|_2^2 + \lambda_1\|\mathbf{W}\|_1\right) \tag{6}$$

We further reduced the dimension for larger CITE-seq datasets by subjecting the feature matrix **P** to singular value decomposition prior to training. Full details and a derivation of the reduced optimization problem are provided elsewhere (16). We fit the elastic-net regression model using the SLEP MATLAB package and evaluated performance with 5-fold cross-validation.

We used the trained **W** to obtain different views of a CITE-seq data set: to infer the TF activities in each cell, we right-multiply the surface protein expression profiles through the model by $\mathbf{WP}^{\mathrm{T}}$; to infer protein activities in each cell, we left-multiply the gene expression profile and TF target-gene interaction matrix through the model by $\mathbf{Y}^{\mathrm{T}}\mathbf{DW}$ (Figure 1C). We refer to these operations as 'mappings' onto the TF space and the surface protein space, respectively.

Significance analysis for TF activities

To assess the statistical significance of the inferred TF activities obtained from the model via the $\mathbf{WP}^{\mathrm{T}}$ mapping, we developed an empirical null model as follows. First, we generated random permutations of the gene expression profiles **Y** for each cell type. For each permuted **Y** response matrix, we trained an AR model using true **D** and **P** input matrices and computed the corresponding inferred TF activities via the $\mathbf{WP}^{\mathrm{T}}$ mapping. Using this permutation and model fitting procedure 5000 times, we generated an empirical null model for activity distribution for each cell. To identify significant TF activities, we assessed the nominal P-value for each cell relative to the empirical null model for the particular regulator TF, and we corrected for multiple hypothesis testing of non-independent hypotheses using the Bonferroni correction procedure. Then, we reported the significant regulators using an adjusted *P*-value of 0.15. We calculated, for each TF regulator, the frequency over samples where the regulator passed its significant threshold for a given cell type. We used this approach to identify significant TF regulators in each cell type to identify the shared and cell type-specific roles TFs.

Pathway analysis

We obtained pathway annotations from MSigDB (19) (c2.all.v7.1.symbols.gmt). This collection is curated from various sources, including online pathway databases (e.g. canonical pathways from BIOCARTA, KEGG, PID, REACTOME and WikiPathways) and the biomedical literature. Using these reference pathway annotations, we constructed a pathway co-occurrence matrix between TFs and surface proteins. For cell-type, we filled a two-way contingency table, with rows representing TF-surface protein pairs that are present in at least one pathway or absent in all pathways based on our pathway co-occurrence matrix, and the columns representing TF-surface protein pairs that are correlated (absolute value of correlation >0.4) or not correlated (absolute value of correlation <0.2). Then we performed enrichment analysis using hypergeometric test based on this contingency table and calculated a *P*-value.

Clustering cells

Hierarchical clustering of surface protein expression and SPaRTAN-predicted TF activities was performed using pvclust (2.2–0) (20), and command pvclust (data, nboot = 1000, method.hclust = 'ward.D2', method.dist = 'correlation'). To identify sub-clusters, we initially split the dendrogram into 10 groups and performed differential protein expression analysis between cells in a given group vs. those in all other groups using Wilcoxon signed rank test for each surface protein and we corrected for multiple hypotheses across surface proteins. We repeated the process and decreased group size until all clusters had at least two differential surface protein (FDR < 0.05) compared to all other clusters.

Running SCENIC

SCENIC (6) is a computational framework that predicts TF activities from scRNA-seq data. We inferred cell-specific TF activities using the SCENIC (R implementation (1.1.2)). We used the cis-regulatory DNA-motif database (hg19-500 bp-upstream-7species.mc9nr.feather, from https://resources.aertslab.org/cistarget/) with default parameters. We computed correlation between SCENIC inferred TF activities and surface protein expression for each cell-type and performed pathway enrichment analysis as outlined above.

Flow Cytometry validation

Single cell suspensions were stained with antibodies against surface proteins (list of ab markers and clone; Supplementary Table S3) for 30 min at 4°C. Dead cells were discriminated by staining with Fixable Viability Dye (eBioscience) in PBS. The cells were washed, and fixed, and permeabilized) for 1 h followed by two wash steps with permeabilization buffer (eBioscience). Then, intracellular staining of transcription factors was conducted for 30 min at 4°C. Flow cytometry analysis was performed by using a Fortessa II (BD Bioscience). Flow cytometric data analyses were performed with FlowJo (Tree Star).

Immunohistochemistry

The population considered for this study consisted of two patients diagnosed with MPeM. These FFPE MPeM surgical specimens were obtained from the National Mesothelioma Virtual Bank (NMVB). The slides are deparaffinized at 60$^{\underline{o}}$C for 30 min and rehydrated using a standard histology protocol. Antigen retrieval was performed using an EDTA buffer (#14747, Cell Signaling, Danvers, MA) in Decloaking chamber at 120°C for 2 min. The slides were stained using an Autostainer Plus (Agilent Dako) platform with TBST rinse buffer (#9997, Cell Signaling). The IHC slides were treated with 3% hydrogen peroxide for 5 min. The primary antibody, BCL-3 (Rabbit Polyclonal, Proteintech Group, Rosemont, IL) was applied using a dilution of 1:100, at room temperature for 45 min. The detection applied, consisted of SignalStain Boost HRP Rabbit (Cell Signaling) for 30 minutes at room temperature. The substrate, 3,3-diaminobenzidine + (# K3468, Agilent Dako), was applied for 8 minutes. The slides were then incubated in Denature solution (Biocare Medical, Pacheco, CA). Following that step, the second primary antibodies, PD-1 (Mouse monoclonal, NAT105, Abcam, Cambridge, MA) was applied at a dilution of 1:50 for 60 min was applied at a dilution of 1:200 for 60 min. The secondary antibodies were followed by Mach 2 Mouse AP (Biocare Medical) detection and Mach 2 Rabbit AP (Biocare Medical), respectively for 45 min. The second chromogen, Warp Red (Biocare Medical) was applied for 10 min. The slides were then counterstained with Hematoxylin (#K8018, Agilent Dako). Digital images of these slides (stained with 2 antibodies) were scanned at 400× magnification on an Aperio AT2 (Leica, Buffalo Grove, IL). After scanning, the third antibody, CD8 (Rabbit monoclonal, SP16, Invitrogen, Carlsbad, CA) was applied using a 1:100 dilution for 60 min at room temp. The detection, Mach 2 Rabbit HRP (Biocare Medical) was applied for 45 min at room temperature. The third chromogen used was Vina Green (Biocare Medical) for 10 min. Digital images of the final stained slides (stained with three antibodies) were similarly generated as before.

Statistical analysis and visualization

Statistical tests were performed with the R 4.0.2 statistical environment. For population comparisons of inferred TF activities and surface protein expression, we performed two-tailed Wilcoxon signed rank test and determined the direction of shifts by comparing the mean of two populations. We corrected raw *P*-values for multiple hypothesis testing based on two methods: Bonferroni and false discovery rate (BH method).

Graphs were generated using RColor-Brewer (version: 1.1 2), ggplot2 (version: 3.3.3) and ComplexHeatmap (version: 2.4.3), ggrepel (version: 0.9.1), circlize (version: 0.4.13) packages. For general data analysis and manipulation, dplyr (version: 1.0.7), matrixStats (version: 0.59.0) and data.table (version: 1.14.0) were used.

RESULTS

SPaRTAN learns a cell-type specific interaction model for cell-surface receptors and transcription factors

SPaRTAN integrates parallel single-cell proteomic and transcriptomic data (based on CITE-seq) with *cis*-regulatory information (e.g. TF:target–gene priors) for predicting cell-specific TF activities and surface protein expression for linking surface receptor signaling to downstream TFs (Figure 1A). Formally, we used an affinity regression (AR) (16) algorithm, a general statistical framework for any problem where the observed data can be explained as interactions between two kinds of inputs, to establish an interaction matrix (**W**) between surface receptors/proteins (**P**) and TFs (**D**) that predicts target gene expression (**Y**) (Figure 1B). To determine the set of TFs that potentially regulate each gene (**D**), we utilized curated TF target-gene interactions (18). We trained independent SPaRTAN models for each cell type that explain gene expression across cells (**Y**) in terms of surface protein expression (**P**) and TF target-gene interactions (**D**) (see Materials and Methods for details).

We use the trained interaction matrix (**W**) to obtain different views of a CITE-seq data set; for example, to predict TF activity from a cell's surface protein expression profile ($\mathbf{WP}^{\mathrm{T}}$) or to predict surface protein expression from a cell's gene expression profile ($\mathbf{Y}^{\mathrm{T}}\mathbf{DW}$) (Figure 1C). Intuitively, information flows down from observed surface protein levels through the learned interaction matrix to infer TF activities and observed mRNA expression levels or propagates up through the TF target-gene edges and interaction network to infer surface protein expression. Importantly, we use these predicted TF activities and surface protein expression to gain biological insights into different cell types and states as described below.

To evaluate our approach, we first trained cell-type specific SPaRTAN models using an existing peripheral blood mononuclear cells (PBMCs) CITE-seq dataset (Supplemental Table S1). Cell types were identified using both protein and gene expression data, the latter with marker genes (see Materials and Methods, Supplementary Figures S1 and S2). For statistical evaluation, we computed the mean Spearman correlation between predicted and measured gene expression profiles on held-out samples using 5-fold cross-validation for each cell-type specific SPaRTAN model using equal numbers of cells. We obtained significantly better performance than a nearest-neighbour approach (P < 0.001, one-sided Wilcoxon signed-rank test) (Supplementary Figure S3A), where the training domain that is most similar to each test example on the basis of surface protein expression is considered the nearest neighbour, and this neighbor's gene expression is used for prediction as shown in Figure 2A. We used Euclidean distance in the surface protein profiles to identify the nearest neighbour. We next evaluated the approach using an independently generated PBMC CITE-seq dataset (see validation PBMC dataset, Supplemental Table S1) and attained similar performance results (Supplementary Figure S3B). We also used the PBMC-trained SPaRTAN cell-type specific models, to infer surface protein expression ($\mathbf{Y}^{\mathrm{T}}\mathbf{DW}$) for each cell type in training and validation CITE-seq PBMC data sets (see Supplemental Table S1). For surface proteins whose Spearman correlations between measured and inferred activities were above 0.3 on training dataset, we found similarly strong correlations between measured and predicted surface protein levels on the validation CITE-seq data (Supplementary Figure S4).

SPaRTAN identifies cell type-specific TFs

Next, we used our approach to predict cell-type specific activities of TFs ($\mathbf{WP}^{\mathrm{T}}$) in PBMCs. To assess the statistical significance of inferred TF activities, we developed an empirical null model based on randomly permuted gene expression profiles for each cell-type (see Materials and Methods). Then, we asked whether the value of individual TF activities for each cell were significantly low or high relative to the corresponding distribution over permuted data. We corrected for multiple hypotheses across TFs and identified significant shared and cell-type specific TFs. Figure 2B shows the fraction of cells per cell-type where each TF was identified as a significant regulator; the representation encompasses the top 8 most prevalent significant TFs for each indicated cell type. Figure 2C and D show the inferred activity distribution of four TFs identified from our analysis: *BACH1*, *IRF8*, *GATA3* and *SPI1* with overlay of CD19, CD56, CD4 and CD11b surface protein expression. These inferred TFs are known regulators of several of the cell types within the PBMCs including *GATA3* (21) for naïve CD4^{+} and CD8^{+} T cells; *SPI1* (*PU1*) (22,23) for B, CD14^{+}/CD16^{+} monocytes and dendritic cells; *BACH1* (24) for B, NK and CD 4 T cells; *PRDM1* (*BLIMP1*) (25) for B, CD16^{+} monocytes and T cells; *IRF8* for monocytes (26). TFs involved in cellular activation and proliferation like *EGR1* and *MYC* were shared across all cell types (Figure 2B). Importantly, in spite of accurately inferring their activities we did not reliably detect mRNAs for many cell-type-specific TFs, given their low levels of transcript expression (Supplementary Figure S5).

SPaRTAN delineates cell type-specific TFs coupled with cell-surface receptors

To explore the associations between inferred TF activities and surface protein expression at a single-cell level, we first computed Pearson correlation coefficients (PCC) between (inferred) TF activity and surface protein expression for each TF-surface protein pair within each cell-type. Figure 3A–C shows the two-way clustering of TFs and proteins by these pairwise PCC in B, CD8^{+} and CD4^{+} T memory cells (see Supplemental Figure S6 for other cell types). We identified several novel as well as known TF-surface protein relationships for each cell type (e.g. EOMES-CD27 (27), STAT6-CD27 (28) STAT5-TIGIT (29), STAT3-ICOS (CD278) (30–32), SMAD3-CD127, STAT1-CD127 (33,34) in CD8^{+} T cells; STAT1/STAT4/STAT5-CD27 (35–37), PRDM1-HLA-DR in B cells (38); SMAD5-ICOS (CD278) (39), STAT5-CD27 (40), MEF2-PD1 (41) in CD4^{+} Memory T cells). We next evaluated the known pathway overlap between TF and surface proteins. Correlated pairs were enriched for known pathways for most cell types ($P < 10^{-3}$ by the hypergeometric test, minimum one overlapping pathways) compared to the SCENIC (6) which uses only single-cell gene expression measurement for identification of cell-type specific TF activities (Supplemental Table S2). Importantly, the analysis suggests that cell-surface receptors can couple with shared and context-dependent downstream transcriptional regulators in different cell types (Figure 3D-F). For example, SPaRTAN-predicted *FOSL2* activity was correlated with CD27 (member of the TNF receptor super family (42)) protein expression in B, CD8^{+} and CD4^{+} memory T cells. Recent studies have shown that *FOSL2* represses Treg development and controls autoimmunity (43) and can also control autoreactive B cells in patients with Systemic lupus erythematosus (SLE) (44). Moreover, SPaRTAN-predicted *STAT5A* activity was highly correlated with CD27 protein expression in B and CD4^{+} memory T cells and with TIGIT (inhibitory immunoreceptor targeted in antitumor immunotherapy (45)) protein expression in CD8^{+} T cells.

To directly test SPaRTAN predicted relationships between cell surface signaling proteins and TF activities in specific cell types, we performed flow cytometry analysis for select surface receptor-intracellular TF pairs using PBMCs from healthy donors (Figure 3G–I, Supplementary

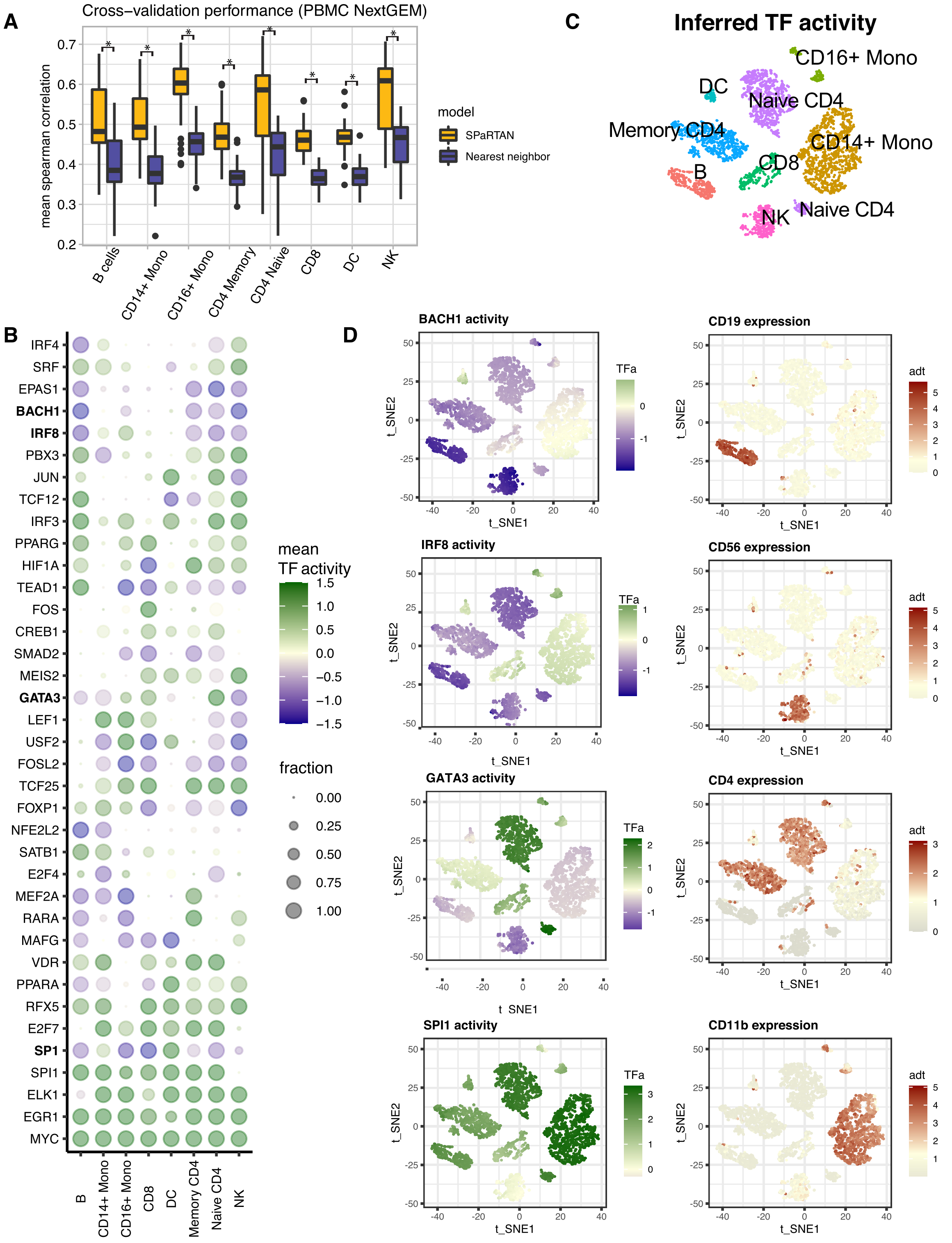


Figure 2. SPaRTAN identifies cell-type specific TFs in peripheral blood mononuclear cells (PBMC). (**A**) SPaRTAN accurately predicts relative gene expression on held-out PBMC (10× Genomics, Next GEM, training dataset) cells for each cell-type. Performance of the SPaRTAN models for each PBMC cell type compared to nearest neighbor method. Boxplots showing mean Spearman correlations between predicted and actual gene expression using the SPaRTAN model (light blue); nearest neighbor by surface protein expression profile (blue) (*y*-axis) for PBMC CITE-seq data from 10× Genomics (Next GEM) each cell-type ($P < 0.001$, one-sided Wilcoxon signed-rank test). (**B**) Dot plot showing the median TF activity z-score of TFs across different cell types. The dot size indicates a fraction of cells for which indicated TF is identified as a significant regulator within designated cell type. For clarity, the union of the top 8 most prevalent significant TFs in each cell type-specific model is shown. (**C**) t-SNE on the inferred TF activity matrix. Cells are colored according to major cell types. (**D**) *BACH1*, *IRF8*, *GATA3*, and *SPI1* inferred TF activity and CD19, CD56, CD4 and CD11b protein expression overlay on t-SNE of TF activities.

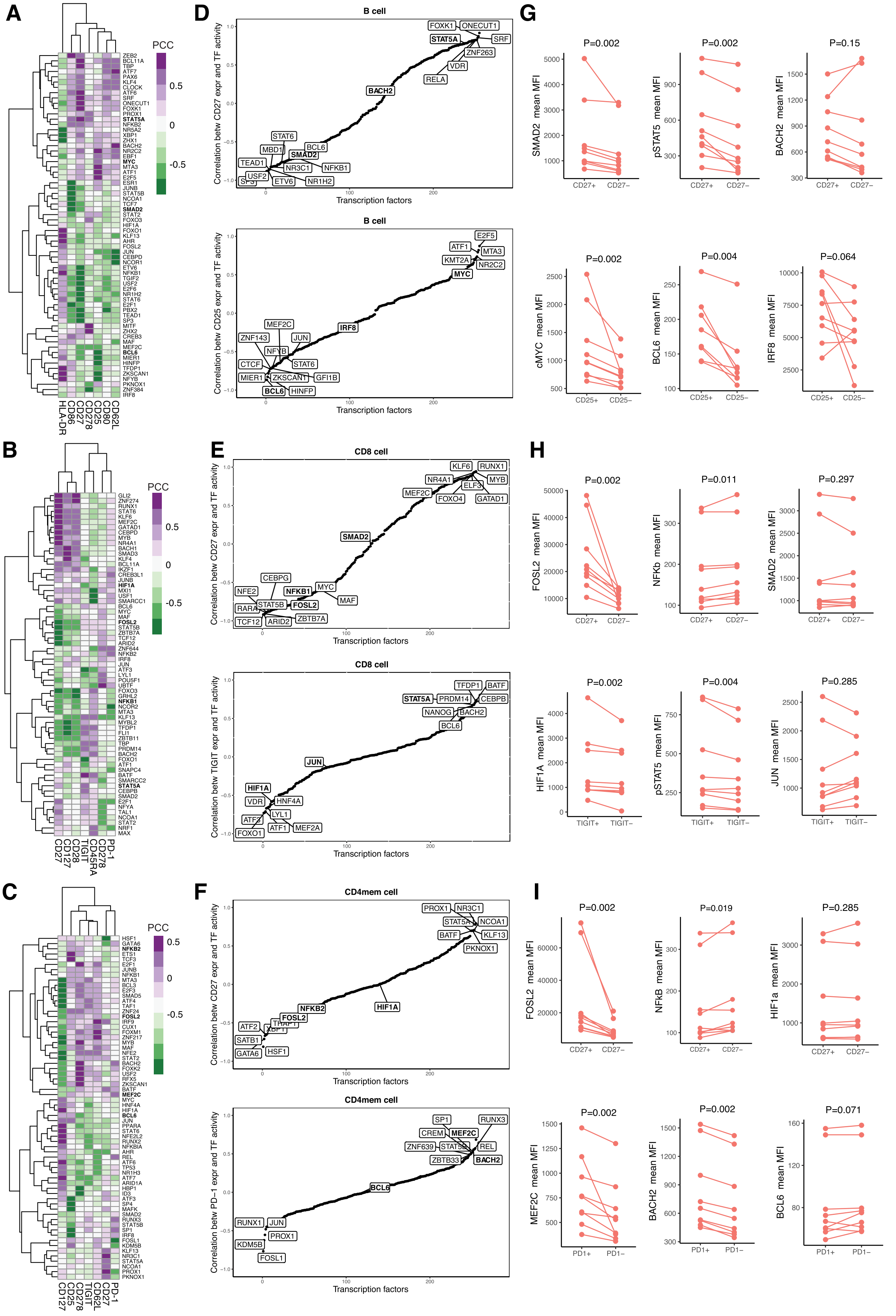


Figure 3. Analysis of SPaRTAN inferred TF activities with cell surface receptor expression in PBMCs—validation with multiparameter flow cytometry. Heatmap revealing correlations between inferred TF activities (rows) and surface protein expression (columns) in (**A**) B cells, (**B**) CD8^{+} T cells and (**C**)

Figure S7). We selected TFs for testing of predictions based on prior knowledge of their functional importance in a given cellular context and the availability of validated antibodies for flow cytometry. Consistent with our predictions and in spite of the variation among individual donors, we observed increased expression of *FOSL2* (transcriptional repressor) and decreased expression of *NFKB* in CD27^{+} CD8^{+} T and CD4^{+} memory T cell subsets compared with their CD27^{-} counterparts; increased expression of *STAT5*, *SMAD2* in CD27^{+} B cells compared with their CD27^{-} counterparts; increased expression of *MYC* and *BCL6* in CD25^{+} (IL-2 receptor) B cells compared with compared with their CD27^{-} counterparts; increased expression of *STAT5* and *HIF1A* in TIGIT^{+} CD8^{+} T cells compared with their TIGIT^{-} counterparts, and increased expression of *MEF2C* and *BACH2* in PD-1^{+} (inhibitory immunoreceptor (46)) CD4^{+} memory T cells compared with their PD-1^{-} counterparts. Importantly, these experimentally validated relationships were not predicted by SCENIC (6) for identification of cell-type specific TF activities using only single-cell gene expression measurements (Supplementary Figure S8) as well as using only TF mRNA levels (Supplementary Figure S9).

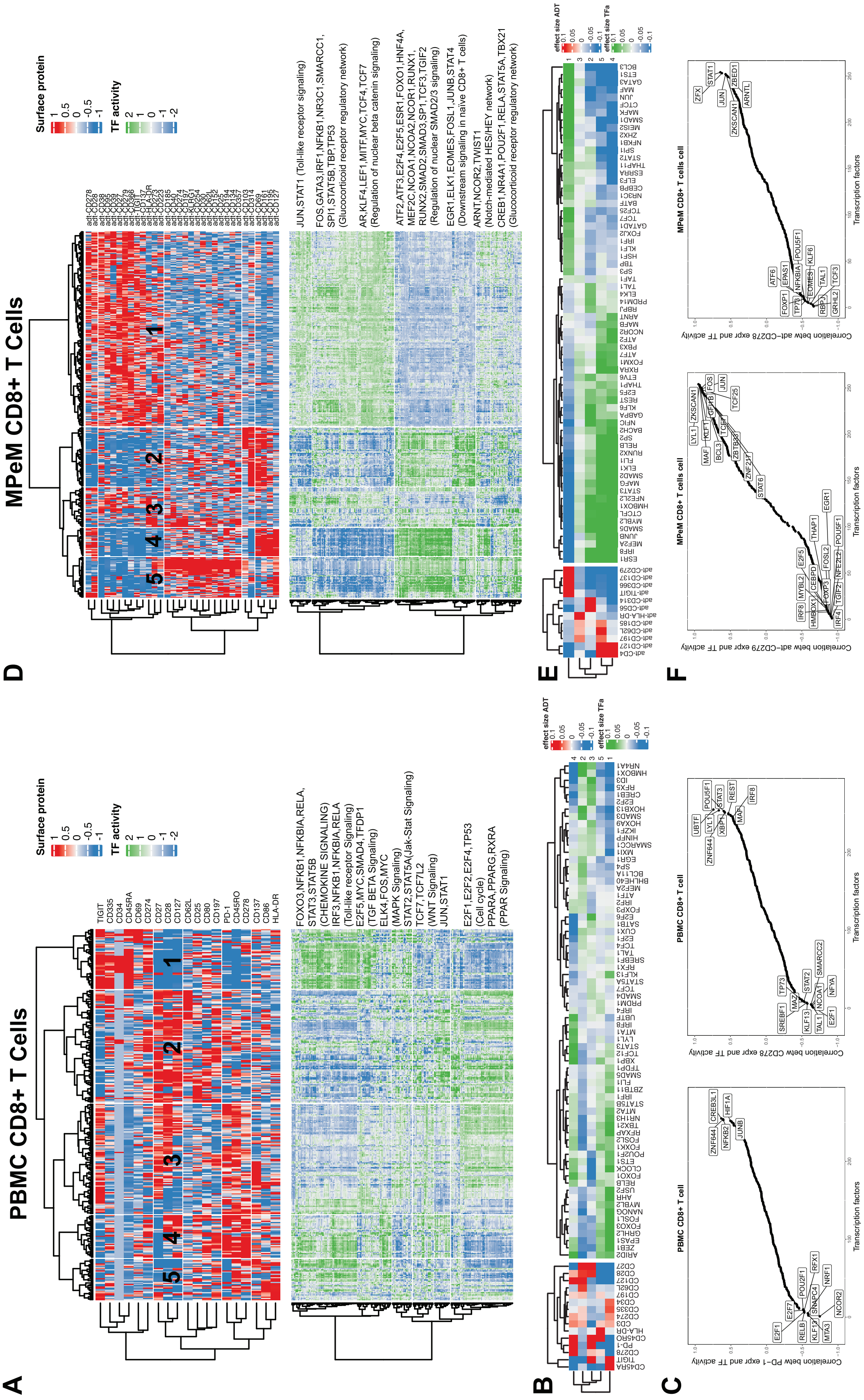
SPaRTAN identifies cell state specific TFs coupled with cell-surface receptors

Next, we asked whether our method could be used to analyze distinct cellular states within a given cell type. Figure 4A shows the clustering of cells by cell surface protein expression (excluding cell lineage marker surface proteins), together with inferred TF activities for the same cell ordering, as derived from the CD8^{+} T cell model. Unsupervised clustering by the surface protein (ADT) expression profiles identified five major cellular states of CD8^{+} T cells. In particular, Cluster 1 was distinguished by high expression for CD45RA and low expression for CD45RO which is characteristic of naïve CD8^{+} T cells (47,48). Among the most significant differences in inferred TF activity associated with this cluster were *KLF13*, *FOSL1*, *USF2*, *FOXO3*, *ZEB1* and *SMAD5* (FDR-corrected $P < 10^{-5}$, t-test; Figure 4B). Cluster 2 was distinguished by cells with high expression of CD27 and CD127 which are characteristic of memory CD8^{+} T cells (49,50). Among the most significant differences in inferred TF activity associated with this cluster were *RUNX1*, *SMAD3*, and *BCL11A* (FDR-corrected $P < 10^{-5}$, *t*-test, Figure 4B).

We performed similar analyses for other cell types (Supplementary Figures S10–S15). For example, unsupervised clustering by surface protein expression identified five major cellular states within the B cell compartment (Supplementary Figure S10). In particular, Cluster 2 was distinguished by B cells with high expression of CD80 which is characteristic of activated B cells (51). Among the most significant differences in inferred TF activity associated with this cluster were *ARID3A*, *TFDP1* and *TP63* (FDR-corrected $P < 10^{-5}$, *t*-test; (Supplementary Figure S10B). Cluster 3 was distinguished by B cells with high expression of CD34 and low expression of CD27 and CD80 which are characteristics of naïve B cells. Among the most significant differences in inferred TF activity associated with this cluster were *TEAD1*, *RFX5*, *JUNB*, *PBX2*, *FOXO3* and *HMBOX1* (FDR-corrected $P < 10^{-5}$, *t*-test (Supplementary Figure S10B). Moreover, Cluster 4 was distinguished by B cells with high expression of CD28 and CD27 and low expression of CD20 which are characteristics of plasmablasts. Among the most significant differences in inferred TF activity associated with this cluster were *MITF*, *STAT5A*, and *FOS* (52–54) (FDR-corrected $P < 10^{-5}$, *t*-test, figure (Supplementary Figure S10B)).

To illustrate the broader utility of our approach, as a discovery platform, we applied it to the tumor microenvironments (TMEs) of malignant peritoneal (MPeM) and pleural (MPM) mesothelioma. Malignant mesothelioma is a rare and aggressive cancer, that has not previously subjected to extensive single-cell profiling and computational analyses. CITE-seq data sets for cells within these TMEs were generated (Supplementary Figures S16 and S17) with a focus on delineating the regulatory states of tumor infiltrating CD8^{+} T cells, given key role in immune surveillance and their manifestation of activated effector or exhausted cell states (55). Unsupervised clustering using cell surface protein expression patterns identified five major populations of MPeM CD8^{+} T cells (Figure 4D). We tested for statistical differences in inferred TF activities and surface protein expression in a given cluster vs. those in all other clusters (Figure 4E). In particular, Cluster 1 was distinguished by MPeM CD8^{+} T cells with high expression for checkpoint inhibitors PD-1, TIM3, and TIGIT which are characteristic of exhausted MPeM CD8^{+} T cells (56). Among the most significant differences in inferred TF activities associated with this cluster were increased values for *TCF7* (57,58), *STAT6*, *BCL3*, and *FOS* (52–54,56) (FDR-corrected $P < 10^{-50}$, *t*-test), some of which have previously been reported as TFs downstream of PD-1 (52–54,56–58) (Figure 4E). Similarly, unsupervised clustering using cell surface protein expression patterns identified five major populations of MPM CD8^{+} T cells (Figure 4G). Cluster 4 was distinguished by MPM CD8^{+} T cells with high expression for checkpoint inhibitors PD-1, TIM3, and TIGIT which is characteristic of exhausted MPM CD8^{+} T cells (56). Similar to MPeM CD8^{+} T cells, among the most sig-

CD4^{+} memory T cells. For clarity, surface proteins with Pearson's correlation coefficient (PCC) values with TFs below 0.75 are filtered, and then the union of the top 10 most correlated TFs with each surface protein is shown for each cell type. Representative sorted correlation plots between (**D**) CD27 (top) CD25 (bottom) protein expression and inferred TF activities across B cells; (**E**) CD27 (top) TIGIT (bottom) protein expression and inferred TF activities across CD8^{+} T cells; (**F**) CD27 (top) PD-1 (bottom) protein expression and inferred TF activities across CD4^{+} memory T cells. (**G–I**) Validation of predictions using flow cytometry analysis. Briefly, B, CD8^{+} and CD4^{+} memory T cells were isolated from peripheral blood (PBL) from healthy donors ($n = 9$) and stained for indicated surface receptors and intracellular TFs. Paired analysis of TF expression was assessed on surface-protein^{+} and surface-protein^{-} cells. P-values are calculated using paired Wilcoxon signed rank test. Representative validation results are shown in each case with two TFs showing high correlation and a control TF showing low correlation.


A
PBMC CD8+ T Cells
Surface protein
TF activity
FOXO3,NFKB1,NFKBIA,RELA,
STAT3,STAT5B
(CHEMOKINE SIGNALING)
IRF3,NFKB1,NFKBIA,RELA
(Toll-like receptor Signaling)
E2F5,MYC,SMAD4,TFDP1
(TGF BETA Signaling)
ELK4,FOS,MYC
(MAPK Signaling)
STAT2,STAT5A(Jak-Stat Signaling)
TCF7,TCF7L2
(WNT Signaling)
JUN,STAT1
E2F1,E2F2,E2F4,TP53
(Cell cycle)
PPARA,PPARG,RXRA
(PPAR Signaling)
B
effect size ADT
effect size TFa
C
PBMC CD8+ T cell
Correlation betw PD−1 expr and TF activity
Correlation betw CD278 expr and TF activity
Transcription factors
D
MPeM CD8+ T Cells
JUN,STAT1 (Toll-like receptor signaling)
FOS,GATA3,IRF1,NFKB1,NR3C1,SMARCC1,
SP1,STAT5B,TBP,TP53
(Glucocorticoid receptor regulatory network)
AR,KLF4,LEF1,MITF,MYC,TCF4,TCF7
(Regulation of nuclear beta catenin signaling)
ATF2,ATF3,E2F4,E2F5,ESR1,FOXO1,HNF4A,
MEF2C,NCOA1,NCOA2,NCOR1,RUNX1,
RUNX2,SMAD2,SMAD3,SP1,TCF3,TGIF2
(Regulation of nuclear SMAD2/3 signaling)
EGR1,ELK1,EOMES,FOSL1,JUNB,STAT4
(Downstream signaling in naive CD8+ T cells)
ARNT,NCOR2,TWIST1
(Notch-mediated HES/HEY network)
CREB1,NR4A1,POU2F1,RELA,STAT5A,TBX21
(Glucocorticoid regulatory network)
E
F
MPeM CD8+ T cells cell
Correlation betw adt−CD279 expr and TF activity
Correlation betw adt−CD278 expr and TF activity
Transcription factors

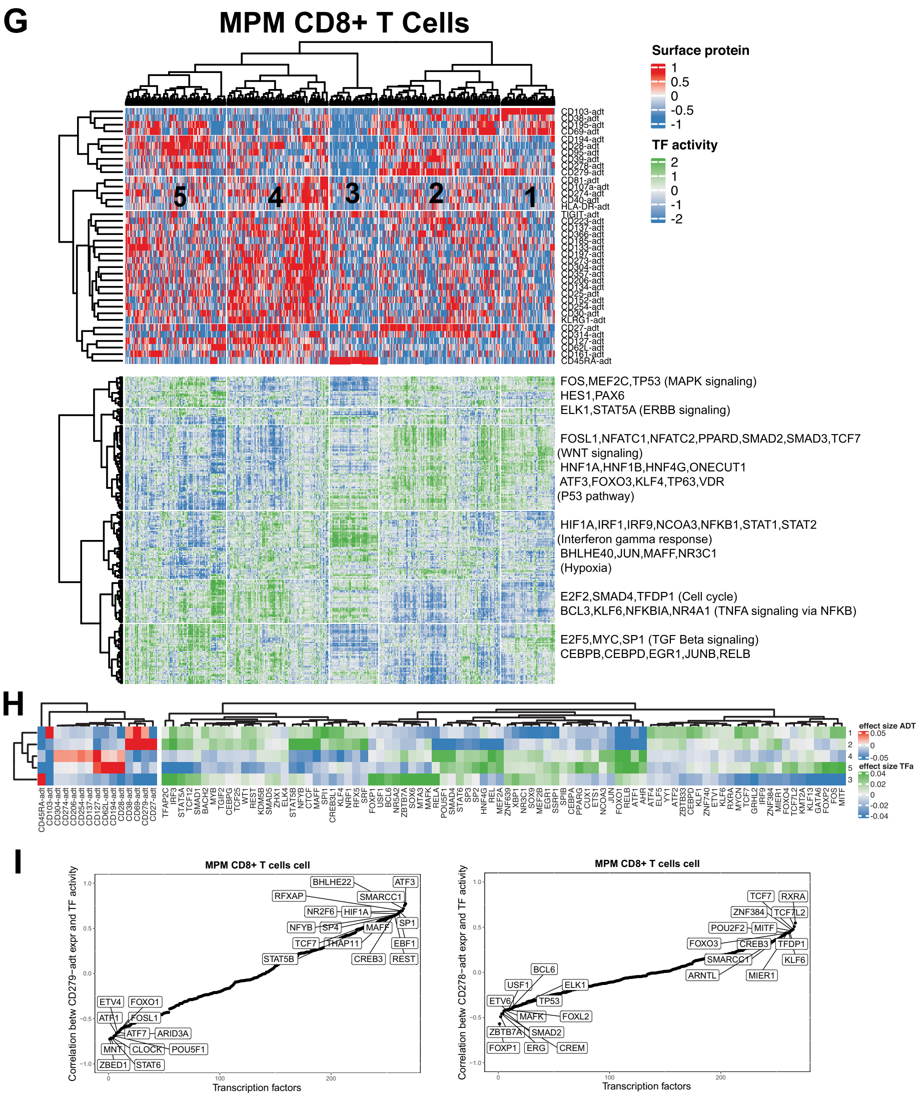


Figure 4. SPaRTAN modeling and analysis of regulatory states of circulating and tumor infiltrating CD8 T cells. (**A**) SPaRTAN model trained on 319 CD8$^+$ T cells from the 10x Genomics PBMC dataset. The top heat map shows cells clustered by the surface protein expression (excluding cell lineage maker surface proteins) The bottom panel shows inferred TF activities for each cell based on clustering by surface protein expression. (**B**) Heatmap shows the mean surface protein expression and inferred TF activity between cells in a given cluster versus those in all other clusters. For each comparison, the absolute value of the mean surface protein expression and inferred TF activity (effect sizes) are ranked and the union of the top 20 TFs for each comparison is shown in the heatmap. (**C**) Sorted correlation plots between PD1 (left) CD278 (ICOS) (right) protein expression and inferred TF activities across PBMC CD8$^+$ T cells. (**D**) We trained a SPaRTAN model on MPeM CD8$^+$ T cells. The top heat map shows cells clustered by the surface protein expression (excluding cell lineage maker surface proteins) The bottom panel shows TF activities for each cell based on clustering by surface protein expression. (**E**) Heatmap shows the mean surface protein expression and inferred TF activity between cells in a given cluster vs. those in all other clusters. For each comparison, the absolute value of the mean surface protein expression and inferred TF activity (effect sizes) are ranked and the union of the top 20 TFs for each comparison is shown in the heatmap. (**F**) Sorted correlation plots between CD279 (PD1) (left) CD278 (ICOS) (right) protein expression and inferred TF activities across MPeM CD8$^+$ T cells. (**G**) SPaRTAN model trained on MPM CD8$^+$ T cells. The top heat map shows cells clustered by the surface protein expression (excluding cell lineage maker surface proteins) The bottom panel shows TF activities for each cell based on clustering by surface protein expression. (**H**) Heatmap shows the mean surface protein expression and inferred TF activity between cells in a given cluster vs. those in all other clusters. For each comparison, the absolute value of the mean surface protein expression and inferred TF activity (effect sizes) are ranked and the union of the top 20 TFs for each comparison is shown in the heatmap. (**I**) Sorted correlation plots between CD279 (PD1) (left) CD278 (ICOS) (right) protein expression and inferred TF activities across MPM CD8$^+$ T cells.

nificant differences in inferred TF activity associated with this cluster was increased *TCF7* (57,58) (FDR-corrected $P < 10^{-50}$, *t*-test) (Figure 4G and H).

Our analysis suggests that cell-surface receptors including those targeted by checkpoint inhibitors in tumor immunotherapy (e.g. PD-1) or T-cell co-stimulatory molecules (e.g. ICOS (inducible T-cell COStimulator (59)) can couple with common and context-dependent downstream TFs within a given cell type but in different tissue contexts (Figure 4C,F,I). For example, SPaRTAN-predicted *HIF1A* activity was correlated with PD-1 (CD-279) protein expression in PBMC, MePM and MPM CD8$^+$ T cells (60,61). Whereas, SPaRTAN-predicted *CTCF*, *CREB3*, *NR2F6*, *TCF7* (58) and *STAT5B* (62) activities were highly correlated with PD-1 protein expression in MPM and MPeM CD8$^+$ T cells. There were also novel TFs correlated with PD-1 protein expression in CD8$^+$ T cells only in one tissue type, including *FOS*, *MYC*, *RARA*, *CEBPB*, *TBP*, *BCL3*, *SREBF1*, *GATA6*, *KLF5*, *ZHX2*, *TFAP4*, *NR2C2*, *MAF* (63) and *ZNF217* for MPeM; *MAFF*, *THAP11* and *RFXAP* for MPM (Supplementary Figure S18, Table S4).

Most of the identified TFs lack prior reports of a link to PD-1, making them potential candidates for follow-up studies. For example, SPaRTAN-predicted *BCL3* activity was correlated with PD-1 protein expression in MePM CD8$^+$ T cells. *BCL3* induces survival and proliferation in cancer cells (64). However, its role in CD8$^+$ T cells as well as in other immune cells has not been studied. To further explore PD-1 coupling with BCL3 in MPeM CD8$^+$ T cells, we performed immunohistochemistry on MPeM specimens. Indeed, we found co-expression of PD-1 and BCL3 in MPeM CD8$^+$ T cells at the protein level (Figure 5).

DISCUSSION

SPaRTAN is a generally applicable method for exploiting parallel single-cell proteomic and transcriptomic data (based on CITE-seq) with *cis*-regulatory information (e.g. TF–target gene priors) to predict the coupling of TF activities with signaling receptors and pathways. Once SPaRTAN is trained using cell-type specific datasets, we utilize the context-specific models to represent individual cells in terms of surface protein expression and TFs' activities. These representations generate hypotheses focused on signaling regulated TF expression and function that can be testable in the laboratory or in the clinic. Application of SPaRTAN to CITE-seq datasets helps to (i) decipher critical regulators (e.g. TFs, surface receptors) underlying cellular identities (e.g. naïve versus memory T cells); (ii) determine whether given cell types have different or common regulators across tissues (e.g. B cells in spleen versus lung); (iii) determine commonalties as well as differences of cell-specific regulatory programs across healthy individuals and those manifesting a disease.

We used SPaRTAN to delineate surface receptors and TF relationships in various types of immune cells in the blood of healthy individuals. SPaRTAN was also used to analyze CITE-seq datasets generated from malignant peritoneal and pleural mesotheliomas. Malignant mesothelioma is a rare and aggressive cancer, that has not previously been subjected to extensive single-cell profiling and computational analyses. Our combined analyses of immune cells in the blood and the tumor microenvironment suggests that signaling receptors e.g., CD27 and PD-1 can be coupled to common or distinct downstream TFs in different cell types and tissues. For example, SPaRTAN-predicted BCL3 activity was correlated with PD-1 protein expression in MePM CD8$^+$ T cells but not in MPM and PBMC CD8$^+$ T cells. We validated co-expression of PD-1, BCL3 and CD8 in the protein level using independent MPeM patient specimens. Future validation experiments can evaluate the role of PD-1 coupling with BCL3 in MPeM CD8$^+$ T cells. BCL3 expression can be evaluated in the setting of PD-1 modulation in cultured MPeM CD8$^+$ T cells. Ultimately, in vivo validation of the role of PD-1 and BCL3 in CD8$^+$ T cells can be studied using malignant peritoneal mesothelioma mouse models through silencing of PD1 and/or BCL3 and measuring the functional activity of CD8$^+$ T cells.

The method we describe has several limitations. First, our analysis uses curated TF target-gene interactions (18) to determine the set of TFs that potentially regulate each gene. Those interactions were curated and collected from different types of evidence such as literature curated resources, ChIP-seq peaks, TF binding site motifs, and interactions inferred directly from gene expression. Therefore, they are noisy, incomplete, and not context-specific. The SPaRTAN framework can be extended using scATAC-seq or bulk ATAC-seq from sorted cells for more accurate representation in both promoter and enhancer regions as performed in our context of patient-specific predictive regulatory models (65)). Furthermore, we do not represent directionality in the TF–gene interaction matrix (i.e. whether a gene is activated or repressed by a TF). Hence, negative values of inferred TF activities can be meaningfully interpreted by prior knowledge of whether the TF is acting as an activator or as a repressor (e.g. for the case of an activator positive inferred TF activity will correspond to upregulation of its target genes and negative values with downregulation of the target genes. On the other hand, for a TF that is functioning as a repressor, an increase in its positive values will correspond with the downregulation of its target genes whereas increased negative values will translate into upregulation of the targets). Our model currently rests on the assumption that a TF either induces or represses its targets, but some TFs may play either role depending on their coordination with co-factors. These limitations may confound the interpretation of activities of TFs with context-specific activator and repressor roles. We have a fixed gene-target gene representation, where the activity of TFs is inferred by correlation with target expression in a linear model; more complex combinatorics of TF binding are not currently modelled. Thus, cooperatively binding TFs (e.g. AP-1−IRF complexes (66)) which can function in signal integration and combinatorial control of gene expression are not modelled. Additionally, individual signalling receptors may modulate the activities of many TFs, some of which are shared with other receptor systems. Such complexity of receptor-TF signal-transduction crosstalk is not explicitly considered.

SPaRTAN analysis is limited to ∼200 surface proteins (for which CITE-seq-validated barcoded antibodies are commercially available from Biolegend). However, the combination of TFs and surface proteins recovers a broad and extensive array of pathways associated with immune cell states in peripheral blood and in the tumor microenvironment. CITE-seq is currently limited to detect surface-protein and gene expression but antibodies directed against intracellular proteins will be added to future iterations of this system (67,68) and can be easily integrated into our approach. Further, we can identify regulators by querying known pathways for upstream and downstream components of the surface protein - TF axis/connection.

Despite these limitations, SPaRTAN will accelerate the analysis of regulatory states of cells that are controlled by

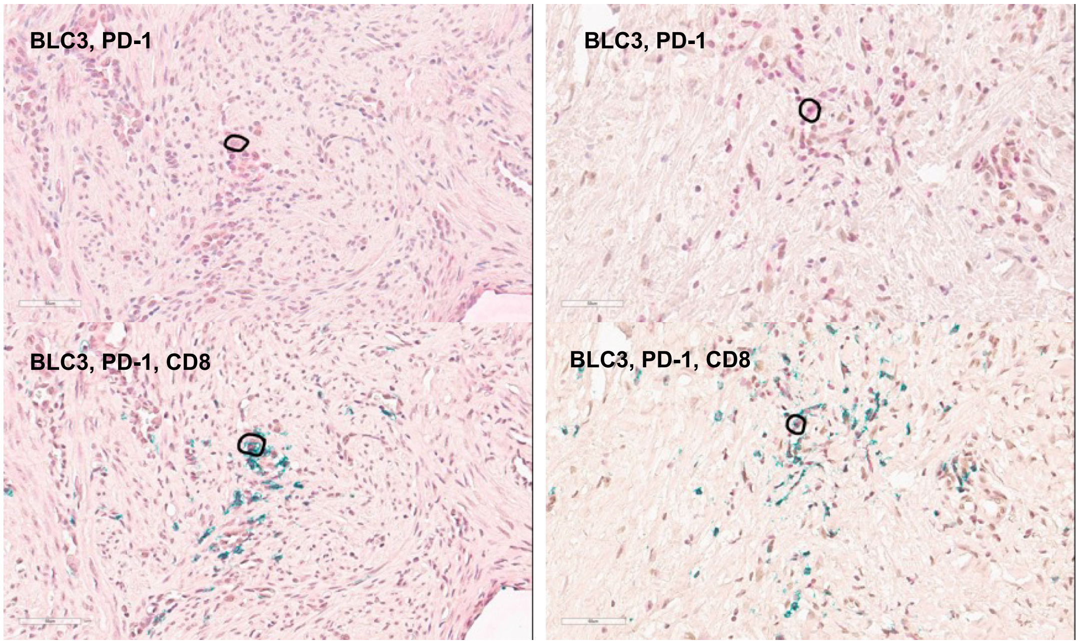


Figure 5. Representative images of immunohistochemistry (IHC) staining of malignant peritoneal mesothelioma tumors using BCL3, PD-1 antibodies (brown and red chromagens, respectively, top) and BCL3, PD-1, CD8 antibodies (brown, red, and green chromagens, respectively, bottom). For illustration doubly (upper panels) or triply (lower panels) stained cells with the indicated antibodies were circled for a cell.

the reciprocal interplay of signaling systems and signal-regulated TFs. It can be used to discover new molecular connections in signal regulated gene expression programs as well as to analyze the cross-talk between signaling pathways. There is substantial variation in immune cell states in healthy individuals depending on age, sex, infection and vaccine history, environmental exposures, diet as well as co-morbidities. Our current analysis is a proof-of-concept but is not large or diverse enough to serve as a reference set. However, extension of our approach with larger datasets can serve to generate a valuable human immune system resource.

DATA AVAILABILITY

The software for SPaRTAN is available from https://github.com/osmanbeyoglulab/SPaRTAN/. Processed data files, inferred TF activities and surface protein and TF activity correlations have been made available the supplementary website for the paper: http://www.pitt.edu/~xim33/SPaRTAN. The published human PBMC CITE-Seq dataset that supports the finding of this study can be downloaded from the 10× Genomics website (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_protein_v3; https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.1.0/5k_pbmc_protein_v3_nextgem). The in-house MPeM and MPM CITE-seq data is available in GEO (https://www.ncbi.nlm.nih.gov/geo/, GSE172155).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Michael J. Becich (PI for the NMVB/Letter of Intent/tissue sample provisioning), Dario Vignali, Tullia Bruno, Anthony Richard Cillo and Feng Shan for helpful discussions, Linan Zhang, Drake Palmer and Devin Dikec for proofreading the manuscript, and Anthony Green for performing tissue staining. We also thank National Mesothelioma Virtual Bank (NMVB) for supplying fresh tumor samples and tissue specimens.

FUNDING

This study was funded by support through the National Institutes of Health (R00 CA207871 to H.U.O.); the Innovation in Cancer Informatics (ICI) Funds (to H.U.O. and A.S.); the UPMC-ITTC fund (to H.S.); the National Institutes of Health (U01 AI141990 to H.S.); CDC NIOSH 2U24OH009077 "National Mesothelioma Virtual Bank, Continued Innovation" (to NMVB and D.J.H). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
2. Peterson,V.M., Zhang,K.X., Kumar,N., Wong,J., Li,L., Wilson,D.C., Moore,R., McClanahan,T.K., Sadekova,S. and Klappenbach,J.A. (2017) Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.*, **35**, 936–939.
3. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.

4. Cusanovich,D.A., Daza,R., Adey,A., Pliner,H.A., Christiansen,L., Gunderson,K.L., Steemers,F.J., Trapnell,C. and Shendure,J. (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–914.
5. Van de Sande,B., Flerin,C., Davie,K., De Waegeneer,M., Hulselmans,G., Aibar,S., Seurinck,R., Saelens,W., Cannoodt,R., Rouchon,Q. *et al.* (2020) A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.*, **15**, 2247–2276.
6. Aibar,S., Gonzalez-Blas,C.B., Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.C., Geurts,P., Aerts,J. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
7. Chan,T.E., Stumpf,M.P.H. and Babtie,A.C. (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.*, **5**, 251–267.
8. Matsumoto,H., Kiryu,H., Furusawa,C., Ko,M.S.H., Ko,S.B.H., Gouda,N., Hayashi,T. and Nikaido,I. (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, **33**, 2314–2321.
9. Fiers,M., Minnoye,L., Aibar,S., Bravo Gonzalez-Blas,C., Kalender Atak,Z. and Aerts,S. (2018) Mapping gene regulatory networks from single-cell omics data. *Brief Funct Genomics*, **17**, 246–254.
10. Schep,A.N., Wu,B., Buenrostro,J.D. and Greenleaf,W.J. (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, **14**, 975–978.
11. Jansen,C., Ramirez,R.N., El-Ali,N.C., Gomez-Cabrero,D., Tegner,J., Merkenschlager,M., Conesa,A. and Mortazavi,A. (2019) Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS Comput. Biol.*, **15**, e1006555.
12. Baron,C.S., Barve,A., Muraro,M.J., van der Linden,R., Dharmadhikari,G., Lyubimova,A., de Koning,E.J.P. and van Oudenaarden,A. (2019) Cell type purification by single-cell transcriptome-trained sorting. *Cell*, **179**, 527–542.
13. Osmanbeyoglu,H.U., Pelossof,R., Bromberg,J.F. and Leslie,C.S. (2014) Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res.*, **24**, 1869–1880.
14. Osmanbeyoglu,H.U., Toska,E., Chan,C., Baselga,J. and Leslie,C.S. (2017) Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat. Commun.*, **8**, 14249.
15. Hmeljak,J., Sanchez-Vega,F., Hoadley,K.A., Shih,J., Stewart,C., Heiman,D., Tarpey,P., Danilova,L., Drill,E., Gibb,E.A. *et al.* (2018) Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discov.*, **8**, 1548–1565.
16. Pelossof,R., Singh,I., Yang,J.L., Weirauch,M.T., Hughes,T.R. and Leslie,C.S. (2015) Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat. Biotechnol.*, **33**, 1242–1249.
17. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
18. Garcia-Alonso,L., Holland,C.H., Ibrahim,M.M., Turei,D. and Saez-Rodriguez,J. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363–1375.
19. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdottir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.
20. Suzuki,R. and Shimodaira,H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.
21. Ho,I.C., Tai,T.S. and Pai,S.Y. (2009) GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. *Nat. Rev. Immunol.*, **9**, 125–135.
22. Oikawa,T., Yamada,T., Kihara-Negishi,F., Yamamoto,H., Kondoh,N., Hitomi,Y. and Hashimoto,Y. (1999) The role of Ets family transcription factor PU.1 in hematopoietic cell differentiation, proliferation and apoptosis. *Cell Death Differ.*, **6**, 599–608.
23. Hosokawa,H., Ungerback,J., Wang,X., Matsumoto,M., Nakayama,K.I., Cohen,S.M., Tanaka,T. and Rothenberg,E.V. (2018) Transcription factor PU.1 represses and activates gene expression in early T cells by redirecting partner transcription factor binding. *Immunity*, **48**, 1119–1134.
24. Igarashi,K., Kurosaki,T. and Roychoudhuri,R. (2017) BACH transcription factors in innate and adaptive immunity. *Nat. Rev. Immunol.*, **17**, 437–450.
25. Nutt,S.L., Fairfax,K.A. and Kallies,A. (2007) BLIMP1 guides the fate of effector B and T cells. *Nat. Rev. Immunol.*, **7**, 923–927.
26. Tamura,T., Kurotaki,D. and Koizumi,S. (2015) Regulation of myelopoiesis by the transcription factor IRF8. *Int. J. Hematol.*, **101**, 342–351.
27. van Aalderen,M.C., Remmerswaal,E.B., Verstegen,N.J., Hombrink,P., ten Brinke,A., Pircher,H., Kootstra,N.A., ten Berge,I.J. and van Lier,R.A. (2015) Infection history determines the differentiation state of human CD8+ T cells. *J. Virol.*, **89**, 5110–5123.
28. Willinger,T., Freeman,T., Hasegawa,H., McMichael,A.J. and Callan,M.F. (2005) Molecular signatures distinguish human central memory from effector memory CD8 T cell subsets. *J. Immunol.*, **175**, 5895–5903.
29. Chauvin,J.M., Ka,M., Pagliano,O., Menna,C., Ding,Q., DeBlasio,R., Sanders,C., Hou,J., Li,X.Y., Ferrone,S. *et al.* (2020) IL15 stimulation with TIGIT blockade reverses CD155-mediated NK-cell dysfunction in melanoma. *Clin. Cancer Res.*, **26**, 5520–5533.
30. Ma,C.S., Avery,D.T., Chan,A., Batten,M., Bustamante,J., Boisson-Dupuis,S., Arkwright,P.D., Kreins,A.Y., Averbuch,D., Engelhard,D. *et al.* (2012) Functional STAT3 deficiency compromises the generation of human T follicular helper cells. *Blood*, **119**, 3997–4008.
31. Tian,Y. and Zajac,A.J. (2016) IL-21 and T cell differentiation: consider the context. *Trends Immunol.*, **37**, 557–568.
32. Doreau,A., Belot,A., Bastid,J., Riche,B., Trescol-Biemont,M.C., Ranchin,B., Fabien,N., Cochat,P., Pouteil-Noble,C., Trolliet,P. *et al.* (2009) Interleukin 17 acts in synergy with B cell-activating factor to influence B cell biology and the pathophysiology of systemic lupus erythematosus. *Nat. Immunol.*, **10**, 778–785.
33. Chakraborty,S., Panda,A.K., Bose,S., Roy,D., Kajal,K., Guha,D. and Sa,G. (2017) Transcriptional regulation of FOXP3 requires integrated activation of both promoter and CNS regions in tumor-induced CD8(+) Treg cells. *Sci. Rep.*, **7**, 1628.
34. Mockus,T.E., Netherby-Winslow,C.S., Atkins,H.M., Lauver,M.D., Jin,G., Ren,H.M. and Lukacher,A.E. (2020) CD8 T cells and STAT1 signaling are essential codeterminants in protection from polyomavirus encephalopathy. *J. Virol.*, **94**, e02038-19.
35. Ueno,H. (2020) The IL-12-STAT4 axis in the pathogenesis of human systemic lupus erythematosus. *Eur. J. Immunol.*, **50**, 10–16.
36. Schmidlin,H., Diehl,S.A. and Blom,B. (2009) New insights into the regulation of human B-cell differentiation. *Trends Immunol.*, **30**, 277–285.
37. Aue,A., Szelinski,F., Weissenberg,S.Y., Wiedemann,A., Rose,T., Lino,A.C. and Dorner,T. (2020) Elevated STAT1 expression but not phosphorylation in lupus B cells correlates with disease activity and increased plasmablast susceptibility. *Rheumatology (Oxford)*, **59**, 3435–3442.
38. Shaffer,A.L., Lin,K.I., Kuo,T.C., Yu,X., Hurt,E.M., Rosenwald,A., Giltnane,J.M., Yang,L., Zhao,H., Calame,K. *et al.* (2002) Blimp-1 orchestrates plasma cell differentiation by extinguishing the mature B cell gene expression program. *Immunity*, **17**, 51–62.
39. Pages,F., Kirilovsky,A., Mlecnik,B., Asslaber,M., Tosolini,M., Bindea,G., Lagorce,C., Wind,P., Marliot,F., Bruneval,P. *et al.* (2009) In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.*, **27**, 5944–5951.
40. Majri,S.S., Fritz,J.M., Villarino,A.V., Zheng,L., Kanellopoulou,C., Chaigne-Delalande,B., Gronholm,J., Niemela,J.E., Afzali,B., Biancalana,M. *et al.* (2018) STAT5B: a differential regulator of the life and death of CD4(+) effector memory T cells. *J. Immunol.*, **200**, 110–118.
41. Di Giorgio,E., Wang,L., Xiong,Y., Akimova,T., Christensen,L.M., Han,R., Samanta,A., Trevisanut,M., Bhatti,T.R., Beier,U.H. *et al.* (2020) MEF2D sustains activation of effector Foxp3+ Tregs during transplant survival and anticancer immunity. *J. Clin. Invest.*, **130**, 6242–6260.
42. Borst,J., Hendriks,J. and Xiao,Y. (2005) CD27 and CD70 in T cell and B cell activation. *Curr. Opin. Immunol.*, **17**, 275–281.
43. Renoux,F., Stellato,M., Haftmann,C., Vogetseder,A., Huang,R., Subramaniam,A., Becker,M.O., Blyszczuk,P., Becher,B., Distler,J.H.W. *et al.* (2020) The AP1 transcription factor Fosl2

promotes systemic autoimmunity and inflammation by repressing Treg development. *Cell Rep.*, **31**, 107826.
44. Scharer,C.D., Blalock,E.L., Mi,T., Barwick,B.G., Jenks,S.A., Deguchi,T., Cashman,K.S., Neary,B.E., Patterson,D.G., Hicks,S.L. *et al.* (2019) Epigenetic programming underpins B cell dysfunction in human SLE. *Nat. Immunol.*, **20**, 1071–1082.
45. Manieri,N.A., Chiang,E.Y. and Grogan,J.L. (2017) TIGIT: a key inhibitor of the cancer immunity cycle. *Trends Immunol.*, **38**, 20–28.
46. Sharpe,A.H. and Pauken,K.E. (2018) The diverse functions of the PD1 inhibitory pathway. *Nat. Rev. Immunol.*, **18**, 153–167.
47. Samji,T. and Khanna,K.M. (2017) Understanding memory CD8(+) T cells. *Immunol. Lett.*, **185**, 32–39.
48. Martin,M.D. and Badovinac,V.P. (2018) Defining memory CD8 T cell. *Front. Immunol.*, **9**, 2692.
49. Hikono,H., Kohlmeier,J.E., Takamura,S., Wittmer,S.T., Roberts,A.D. and Woodland,D.L. (2007) Activation phenotype, rather than central- or effector-memory phenotype, predicts the recall efficacy of memory CD8+ T cells. *J. Exp. Med.*, **204**, 1625–1636.
50. Olson,J.A., McDonald-Hyman,C., Jameson,S.C. and Hamilton,S.E. (2013) Effector-like CD8+ T cells in the memory population mediate potent protective immunity. *Immunity*, **38**, 1250–1260.
51. Good-Jacobson,K.L., Song,E., Anderson,S., Sharpe,A.H. and Shlomchik,M.J. (2012) CD80 expression on B cells regulates murine T follicular helper development, germinal center B cell survival, and plasma cell generation. *J. Immunol.*, **188**, 4217–4225.
52. Xiao,G., Deng,A., Liu,H., Ge,G. and Liu,X. (2012) Activator protein 1 suppresses antitumor T-cell function via the induction of programmed death 1. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 15419–15424.
53. Renkema,K.R., Lee,J.Y., Lee,Y.J., Hamilton,S.E., Hogquist,K.A. and Jameson,S.C. (2016) IL-4 sensitivity shapes the peripheral CD8+ T cell pool and response to infection. *J. Exp. Med.*, **213**, 1319–1329.
54. Stelekati,E., Chen,Z., Manne,S., Kurachi,M., Ali,M.A., Lewy,K., Cai,Z., Nzingha,K., McLane,L.M., Hope,J.L. *et al.* (2018) Long-term persistence of exhausted CD8 T cells in chronic infection is regulated by microRNA-155. *Cell Rep.*, **23**, 2142–2156.
55. Maimela,N.R., Liu,S. and Zhang,Y. (2019) Fates of CD8+ T cells in tumor microenvironment. *Comput. Struct. Biotechnol. J.*, **17**, 1–13.
56. Wherry,E.J. and Kurachi,M. (2015) Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.*, **15**, 486–499.
57. Kurtulus,S., Madi,A., Escobar,G., Klapholz,M., Nyman,J., Christian,E., Pawlak,M., Dionne,D., Xia,J., Rozenblatt-Rosen,O. *et al.* (2019) Checkpoint blockade immunotherapy induces dynamic changes in PD-1(-)CD8(+) tumor-infiltrating T cells. *Immunity*, **50**, 181–194.
58. Siddiqui,I., Schaeuble,K., Chennupati,V., Fuertes Marraco,S.A., Calderon-Copete,S., Pais Ferreira,D., Carmona,S.J., Scarpellino,L., Gfeller,D., Pradervand,S. *et al.* (2019) Intratumoral Tcf1(+)PD-1(+)CD8(+) T cells with stem-like properties promote tumor control in response to vaccination and checkpoint blockade immunotherapy. *Immunity*, **50**, 195–211.
59. Wikenheiser,D.J. and Stumhofer,J.S. (2016) ICOS co-stimulation: friend or foe? *Front. Immunol.*, **7**, 304.
60. Barsoum,I.B., Smallwood,C.A., Siemens,D.R. and Graham,C.H. (2014) A mechanism of hypoxia-mediated escape from adaptive immunity in cancer cells. *Cancer Res.*, **74**, 665–674.
61. Samanta,D., Park,Y., Ni,X., Li,H., Zahnow,C.A., Gabrielson,E., Pan,F. and Semenza,G.L. (2018) Chemotherapy induces enrichment of CD47(+)/CD73(+)/PDL1(+) immune evasive triple-negative breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E1239–E1248.
62. Chen,Y., Yu,M., Zheng,Y., Fu,G., Xin,G., Zhu,W., Luo,L., Burns,R., Li,Q.Z., Dent,A.L. *et al.* (2019) CXCR5(+)PD-1(+) follicular helper CD8 T cells control B cell tolerance. *Nat. Commun.*, **10**, 4415.
63. Andris,F., Denanglaire,S., Anciaux,M., Hercor,M., Hussein,H. and Leo,O. (2017) The transcription factor c-Maf promotes the differentiation of follicular helper T cells. *Front. Immunol.*, **8**, 480.
64. Maldonado,V. and Melendez-Zajgla,J. (2011) Role of Bcl-3 in solid tumors. *Mol. Cancer*, **10**, 152.
65. Osmanbeyoglu,H.U., Shimizu,F., Rynne-Vidal,A., Alonso-Curbelo,D., Chen,H.A., Wen,H.Y., Yeung,T.L., Jelinic,P., Razavi,P., Lowe,S.W. *et al.* (2019) Chromatin-informed inference of transcriptional programs in gynecologic and basal breast cancers. *Nat. Commun.*, **10**, 4369.
66. Glasmacher,E., Agrawal,S., Chang,A.B., Murphy,T.L., Zeng,W., Vander Lugt,B., Khan,A.A., Ciofani,M., Spooner,C.J., Rutz,S. *et al.* (2012) A genomic regulatory element that directs assembly and function of immune-specific AP-1-IRF complexes. *Science*, **338**, 975–980.
67. Mimitou,E.P., Cheng,A., Montalbano,A., Hao,S., Stoeckius,M., Legut,M., Roush,T., Herrera,A., Papalexi,E., Ouyang,Z. *et al.* (2019) Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods*, **16**, 409–412.
68. Chung,H., Parkhurst,C.N., Magee,E.M., Phillips,D., Habibi,E., Chen,F., Yeung,B., Waldman,J., Artis,D. and Regev,A. (2021) Simultaneous single cell measurements of intranuclear proteins and gene expression. bioRxiv doi: https://doi.org/10.1101/2021.01.18.427139, 19 January 2021, preprint: not peer reviewed.