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

Xiaojun Ma[1*], Ashwin Somasundaram[2*], Zengbiao Qi[1], Harinder Singh[3], Hatice Ulku Osmanbeyoglu[1,4+]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department of Medicine, Division of Hematology/Oncology, University of Pittsburgh, Pittsburgh, PA, USA

[3]Center for Systems Immunology and Department of Immunology, University of Pittsburgh, Pittsburgh, PA, USA

[4]Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA

*These authors contributed equally

[+]Corresponding author, email: osmanbeyogluhu@pitt.edu, phone: 412-623-7789

The developmental pathways and functions of specialized cell types are dependent on the complex interplay between signaling and transcriptional networks. 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 peripheral blood mononuclear cells (PBMCs) to predict the coupling of signaling receptors with cell context-specific TF activities. The predictions are validated by flow cytometry analyses. SPaRTAN is then used to analyze the signaling coupled TF activity states of tumor infiltrating CD8$^+$ T cells in malignant peritoneal and pleural mesotheliomas. SPaRTAN greatly enhances the utility of CITE-seq datasets to probe signaling coupled TF networks that regulate developmental or functional transitions in cellular states.

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]. To date, the datasets generated by this powerful platform have not

been used to link the expression of cell surface proteins, e.g., 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 and corresponding scRNA-seq datasets, both obtained using CITE-seq, to link expression of surface proteins with inferred TF activities. Our new 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]. Briefly, our model views expression of surface proteins (antibody-derived tags (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 (**Fig. 1A**). Specifically, we use a regularized bilinear regression algorithm called affinity regression (AR)[16], a general statistical framework for any problem where the observed data can be explained as interactions (**W**) between two kinds of inputs, to establish an interaction matrix between surface proteins (**P**) and TFs (**D**) that predicts target gene expression (**Y**) (**Fig. 1B**). We used curated TF target-gene interactions[17] to determine the set of TFs that potentially regulate each gene. Here, we first trained independent SPaRTAN models for each cell type that explain gene expression across cells in terms of surface protein expression and TF target-gene interactions (see **Methods section**). The model captures statistical relationships between surface proteins, TFs, and gene expression. We can use the trained interaction matrix to obtain different views of a CITE-seq data set; *e.g.*, to predict TF activity from a cell's surface protein expression profile or to predict surface protein expression from a cell's gene expression profile (**Fig. 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. 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.

We trained cell-type specific SPaRTAN models using an existing peripheral blood mononuclear cells (PBMCs) CITE-seq dataset (**Supplemental Table 1**). Cell types were identified using both protein and gene expression data, the latter with marker genes (see **Methods section**, **Supplementary Fig. 1-2**). 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-neighbor approach based on Euclidean distance in the surface protein profiles (input space) (P<0.001, one-sided Wilcoxon's signed-rank test) as shown in **Fig. 1D**. We next evaluated the approach using an independently generated PBMC CITE-seq dataset (see validation PBMC dataset, **Supplemental Table 1**) and attained similar performance results (**Supplementary Fig. 3**). We also used the PBMC-trained SPaRTAN cell-type specific models, to infer surface protein expression ($\mathbf{Y}^T\mathbf{DW}$) for each cell type in training and validation CITE-seq PBMC data sets (see **Supplemental Table 1**). 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 Fig. 4**).

Next, we used our approach to predict cell-type specific activities of TFs ($\mathbf{WP}^T$) 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 **Methods section**). 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**. Fig. 1E** 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. **Fig. 1F-G** show the inferred activity distribution of four TFs identified from our analysis: *BACH1*, *GATA3*, *SPI1*, and *IRF8* 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*[18] for naïve $CD4^+$ and $CD8^+$ T cells; *SPI1* (*PU1*)[19,20] for B, $CD14^+$/$CD16^+$ monocytes and dendritic cells; *BACH1*[21] for B, NK, and CD 4 T cells; *PRDM1* (*BLIMP1*)[22] for B, $CD16^+$ monocytes and T cells. TFs involved in cellular activation and proliferation like *EGR1* and *MYC* were shared across all cell types (**Fig. 1E**). 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 Fig. 5**).

To explore the associations between inferred TF activities and surface protein expression at a single-cell level, we first computed Pearson correlations between (inferred) TF activity and surface protein expression for each TF-surface protein pair within each cell-type. **Fig. 2A-C** (see also **Supplemental Fig. 6**) shows the two-way clustering of TFs and proteins by these pairwise Pearson correlations in B, $CD8^+$ and $CD4^+$ T memory cells. We identified several known as well as novel TF-surface protein relationships for each cell type. 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 alternative approach[6] for identification of cell-type specific TF activities using only single-cell gene expression measurement (**Supplemental Table 2).** Importantly, the analysis suggests that a surface signaling protein can couple with shared and context-dependent downstream transcriptional regulators in different cell types (**Fig. 2D-F**). For example, SPaRTAN-predicted FOSL2 activity was correlated with CD27 (member of the TNF receptor super family[23])

protein expression in B, $CD8^+$ and $CD4^+$ memory T cells. In contrast, 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[24]) 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 (**Fig. 2G-I, Supplementary Fig. 7, Supplementary Table 3-8)**. Consistent with our predictions and in spite of the variation among individual donors, we observed increased expression of FOSL2 in $CD27^+$ $CD8^+$ T and $CD4^+$ memory T cell subsets compared with their $CD27^-$ counterparts; increased expression of STAT5 in $CD25^+$ (IL-2 receptor) B cells compared with the $CD25^-$ subset and increased expression of MEF2C and BACH2 with the expression of PD-1 (inhibitory immunoreceptor[25]) in $CD4^+$ memory T cells. Importantly, these experimentally validated relationships were not predicted by the alternative approach[6] for identification of cell-type specific TF activities using only single-cell gene expression measurements (**Supplementary Fig. 8)** as well as using only TF mRNA levels (**Supplementary Fig. 9).**

Next, we asked whether our method could be used to analyze distinct cellular states within a given cell type. **Fig. 3A** 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. 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; **Fig. 3B**). Cluster 2 was distinguished by cells with high expression of CD27 and CD127 which are characteristic of memory $CD8^+$ T cells. 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, **Fig. 3B**).

We performed similar analyses for other cell types (**Supplementary Fig. 10–15**). For example, unsupervised clustering by surface protein expression identified five major cellular states within the B cell compartment (**Supplementary Fig. 10)**. In particular, Cluster 2 was distinguished by B cells with high expression of CD80 which is characteristic of activated B cells. 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 Fig. 10B**). 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 Fig. 10B**). 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*[26-28] (FDR-corrected $p < 10^{-5}$, t-test, **Fig.** (**Supplementary Fig. 10B**).

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. CITE-seq data sets for cells within these TMEs were generated (**Supplementary Fig. 16-17**) and queried for the regulatory states of tumor infiltrating $CD8^+$ T cells. Unsupervised clustering using

cell surface protein expression patterns identified five major populations of MPeM CD8$^+$ T cells (**Fig. 3D**). We tested for statistical differences in inferred TF activities and surface protein expression in a given cluster vs. those in all other clusters (**Fig. 3E**). 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[29]. Among the most significant differences in inferred TF activities associated with this cluster were increased values for *TCF7*[30,31]*, STAT6, BCL3, and FOS*[26-29] (FDR-corrected $p < 10^{-50}$, t-test), some of which have previously been reported as TFs downstream of PD-1[26-31] (**Fig. 3E**). Similarly, unsupervised clustering using cell surface protein expression patterns identified five major populations of MPM CD8$^+$ T cells (**Fig. 3G**). 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[29]. Similar to MPeM CD8$^+$ T cells, among the most significant differences in inferred TF activity associated with this cluster was increased *TCF7*[30,31] (FDR-corrected $p < 10^{-50}$, t-test) (**Fig. 3G-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[32]) can couple with common and context-dependent downstream TFs within a given cell type but in different tissue contexts (**Fig. 3C,F,I**).

Thus, 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. In the future, the framework can be extended to include single-cell epigenomic data so as to enable the integration of cell signaling with alterations in TF activities, chromatin structure and gene expression. SPaRTAN will accelerate the analysis of regulatory states of cells that are controlled by 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.

Methods

Data and preprocessing. To construct the TF – target gene prior matrix, we downloaded a gene set resource containing TF target-gene interactions from DoRothEA[17]. 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 defines (**D**) a candidate set of associations between TFs and target genes.

CITE-seq data for 5k (Chemistry v3), 5k (Nextgen) PBMC obtained from 10× Genomics website (**Supplementary Table 1**). A total of ~5,000 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 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.

Tissue Processing of Malignant Mesothelioma. Tumor samples were washed in RPMI containing antibiotics such as amphotericin B and penicillin-streptomycin for 30 minutes followed by mechanical and enzymatic digestion and further passage via a 100µm filter. Isolated tumor infiltrating lymphocytes (TIL) were then washed in RPMI media twice, and stained with aformentioned 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 resuspended in PBS (0.04% BSA; Sigma) and then loaded into the 10X Controller for droplet generation, targeting recovery of 5,000 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 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 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 aNextSeq500/550 with a depth of 50K reads per cell.

Raw sequence data were processed via CellRanger 3.0 (10X 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. Antibody counts for CITE-seq were counted using CITE-seq-count. Normalization and initial explanatory analysis of CITE-seq data were performed using the Seurat R package[33]. Antibody-derived tags (ADTs) for each cell will be normalized using a centered log ratio (CLR) transformation across cells. During quality control, we excluded cells with > 5000 and <300 expressed genes. 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.

Training cell-type specific SPaRTAN models. We trained cell-specific SPaRTAN models using affinity regression (AR) which is an algorithm for efficiently solving a regularized bilinear regression problem[13,16], defined here as follows. For a given set of genes and their expression profiles, we will use AR to learn an interaction matrix between cell surface proteins and TFs that explains target gene expression in cells. For a data set of *M* cells profiled using scRNA-seq across *N* genes, we let Y $\in R^{NxM}$ be the log normalized gene expression profiles for all cells, where each column of Y corresponds to a scRNA-seq experiment. We define a matrix D $\in R^{NxQ}$, where each row represents a gene and each column is a binary vector representing the target genes of a TF. Finally, we define a matrix P $\in R^{MxS}$ of cell surface protein attributes where each row represents a cell and each column represents protein expression levels of a surface protein across cells from scADT-seq. We set up a bilinear regression problem to learn an interaction matrix W $\in R^{QxS}$ for pairs of TF-surface protein features that predicts target gene expression.

$$\mathbf{Y} = \mathbf{DWP}^{T}$$

To reduce the dimensionality, we subjected the feature matrix P to singular value decomposition before training and reduced to a smaller system of equations where the output is the set of

pairwise similarities $\mathbf{Y}^T\mathbf{Y}$ between examples rather than **Y** itself. Then we used elastic-net regression to solve for the interaction matrix. Full details and a derivation of the reduced optimization problem are provided elsewhere[16]. Since the model captures relationships between surface proteins, TFs, and gene expression, we can use the trained **W** to obtain different views of a single cell data set: to predict the TF activities in each cell, we can right-multiply the surface protein expression profiles through the model by $\mathbf{WP}^T$; to predict surface protein activities in each cell, we can left-multiply the gene expression profile and motif-hit matrix through the model by $\mathbf{Y}^T\mathbf{DW}$. 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 $\mathbf{WP}^T$ 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 $\mathbf{WP}^T$ 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.

Flow Cytometry validation. Single cell suspensions were stained with antibodies against surface proteins (list of ab markers and clone; **Supplementary Table 9**) for 30 minutes 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 hour followed by two wash steps with permeabilization buffer (eBioscience). Then, intracellular staining of transcription factors were conducted for 30 minutes 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).

Statistical analysis. Statistical tests were performed with the R statistical environment. For population comparisons of inferred TF activities and surface protein expression, we performed two-tailed Wilcoxon's signed-rank tests and determined the direction of shifts by comparing the mean of two populations.

Data availability

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/, GSExx).

Code availability

The software for SPaRTAN is available from https://github.com/osmanbeyoglulab/SPaRTAN

References

1 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

2 Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936-939, doi:10.1038/nbt.3973 (2017).

3 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).

4 Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914, doi:10.1126/science.aab1601 (2015).

5 Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat Protoc* **15**, 2247-2276, doi:10.1038/s41596-020-0336-2 (2020).

6 Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086, doi:10.1038/nmeth.4463 (2017).

7 Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5**, 251-267 e253, doi:10.1016/j.cels.2017.08.014 (2017).

8 Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314-2321, doi:10.1093/bioinformatics/btx194 (2017).

9 Fiers, M. *et al.* Mapping gene regulatory networks from single-cell omics data. *Brief Funct Genomics* **17**, 246-254, doi:10.1093/bfgp/elx046 (2018).

10 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978, doi:10.1038/nmeth.4401 (2017).

11 Jansen, C. *et al.* Building gene regulatory networks from scATAC-seq and scRNA-seq using Linked Self Organizing Maps. *PLoS Comput Biol* **15**, e1006555, doi:10.1371/journal.pcbi.1006555 (2019).

12 Baron, C. S. *et al.* Cell Type Purification by Single-Cell Transcriptome-Trained Sorting. *Cell* **179**, 527-542 e519, doi:10.1016/j.cell.2019.08.006 (2019).

13 Osmanbeyoglu, H. U., Pelossof, R., Bromberg, J. F. & Leslie, C. S. Linking signaling pathways to transcriptional programs in breast cancer. *Genome Res* **24**, 1869-1880, doi:10.1101/gr.173039.114 (2014).

14 Osmanbeyoglu, H. U., Toska, E., Chan, C., Baselga, J. & Leslie, C. S. Pancancer modelling predicts the context-specific impact of somatic mutations on transcriptional programs. *Nat Commun* **8**, 14249, doi:10.1038/ncomms14249 (2017).

15 Hmeljak, J. *et al.* Integrative Molecular Characterization of Malignant Pleural Mesothelioma. *Cancer Discov* **8**, 1548-1565, doi:10.1158/2159-8290.CD-18-0804 (2018).

16 Pelossof, R. *et al.* Affinity regression predicts the recognition code of nucleic acid-binding proteins. *Nat Biotechnol* **33**, 1242-1249, doi:10.1038/nbt.3343 (2015).

17 Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res* **29**, 1363-1375, doi:10.1101/gr.240663.118 (2019).

18 Ho, I. C., Tai, T. S. & Pai, S. Y. GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. *Nat Rev Immunol* **9**, 125-135, doi:10.1038/nri2476 (2009).

19 Oikawa, T. *et al.* The role of Ets family transcription factor PU.1 in hematopoietic cell differentiation, proliferation and apoptosis. *Cell Death Differ* **6**, 599-608, doi:10.1038/sj.cdd.4400534 (1999).

20 Hosokawa, H. *et al.* Transcription Factor PU.1 Represses and Activates Gene Expression in Early T Cells by Redirecting Partner Transcription Factor Binding. *Immunity* **48**, 1119-1134 e1117, doi:10.1016/j.immuni.2018.04.024 (2018).

21 Igarashi, K., Kurosaki, T. & Roychoudhuri, R. BACH transcription factors in innate and adaptive immunity. *Nat Rev Immunol* **17**, 437-450, doi:10.1038/nri.2017.26 (2017).

22 Nutt, S. L., Fairfax, K. A. & Kallies, A. BLIMP1 guides the fate of effector B and T cells. *Nat Rev Immunol* **7**, 923-927, doi:10.1038/nri2204 (2007).

23 Borst, J., Hendriks, J. & Xiao, Y. CD27 and CD70 in T cell and B cell activation. *Curr Opin Immunol* **17**, 275-281, doi:10.1016/j.coi.2005.04.004 (2005).

24 Manieri, N. A., Chiang, E. Y. & Grogan, J. L. TIGIT: A Key Inhibitor of the Cancer Immunity Cycle. *Trends Immunol* **38**, 20-28, doi:10.1016/j.it.2016.10.002 (2017).

25 Sharpe, A. H. & Pauken, K. E. The diverse functions of the PD1 inhibitory pathway. *Nat Rev Immunol* **18**, 153-167, doi:10.1038/nri.2017.108 (2018).

26 Xiao, G., Deng, A., Liu, H., Ge, G. & Liu, X. 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, doi:10.1073/pnas.1206370109 (2012).

27 Renkema, K. R. *et al.* IL-4 sensitivity shapes the peripheral CD8+ T cell pool and response to infection. *J Exp Med* **213**, 1319-1329, doi:10.1084/jem.20151359 (2016).

28 Stelekati, E. *et al.* Long-Term Persistence of Exhausted CD8 T Cells in Chronic Infection Is Regulated by MicroRNA-155. *Cell Rep* **23**, 2142-2156, doi:10.1016/j.celrep.2018.04.038 (2018).

29 Wherry, E. J. & Kurachi, M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol* **15**, 486-499, doi:10.1038/nri3862 (2015).

30 Kurtulus, S. *et al.* Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD-1(-)CD8(+) Tumor-Infiltrating T Cells. *Immunity* **50**, 181-194 e186, doi:10.1016/j.immuni.2018.11.014 (2019).

31 Siddiqui, I. *et al.* 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 e110, doi:10.1016/j.immuni.2018.12.021 (2019).

32 Wikenheiser, D. J. & Stumhofer, J. S. ICOS Co-Stimulation: Friend or Foe? *Front Immunol* **7**, 304, doi:10.3389/fimmu.2016.00304 (2016).

33 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502, doi:10.1038/nbt.3192 (2015).

Acknowledgments

This work was supported by awards from Innovation in Cancer Informatics (ICI) Funds and R00 CA207871. We thank Dario Vignali, Tullia Bruno, Anthony Richard Cillo and Feng Shan for helpful discussions. We also thank National Mesothelioma Virtual Bank (NMVB) for supplying fresh tumor samples.

Author contributions

H.U.O. designed the study, developed algorithmic approaches, performed all computational experiments and analyses and wrote the manuscript. X.M. implemented python version of SPaRTAN. A.S. and Z.Q. performed the experimental validation and helped to analyze results and write the experimental validation section. H. S. helped to analyze results, structure and edit the manuscript.

Figures

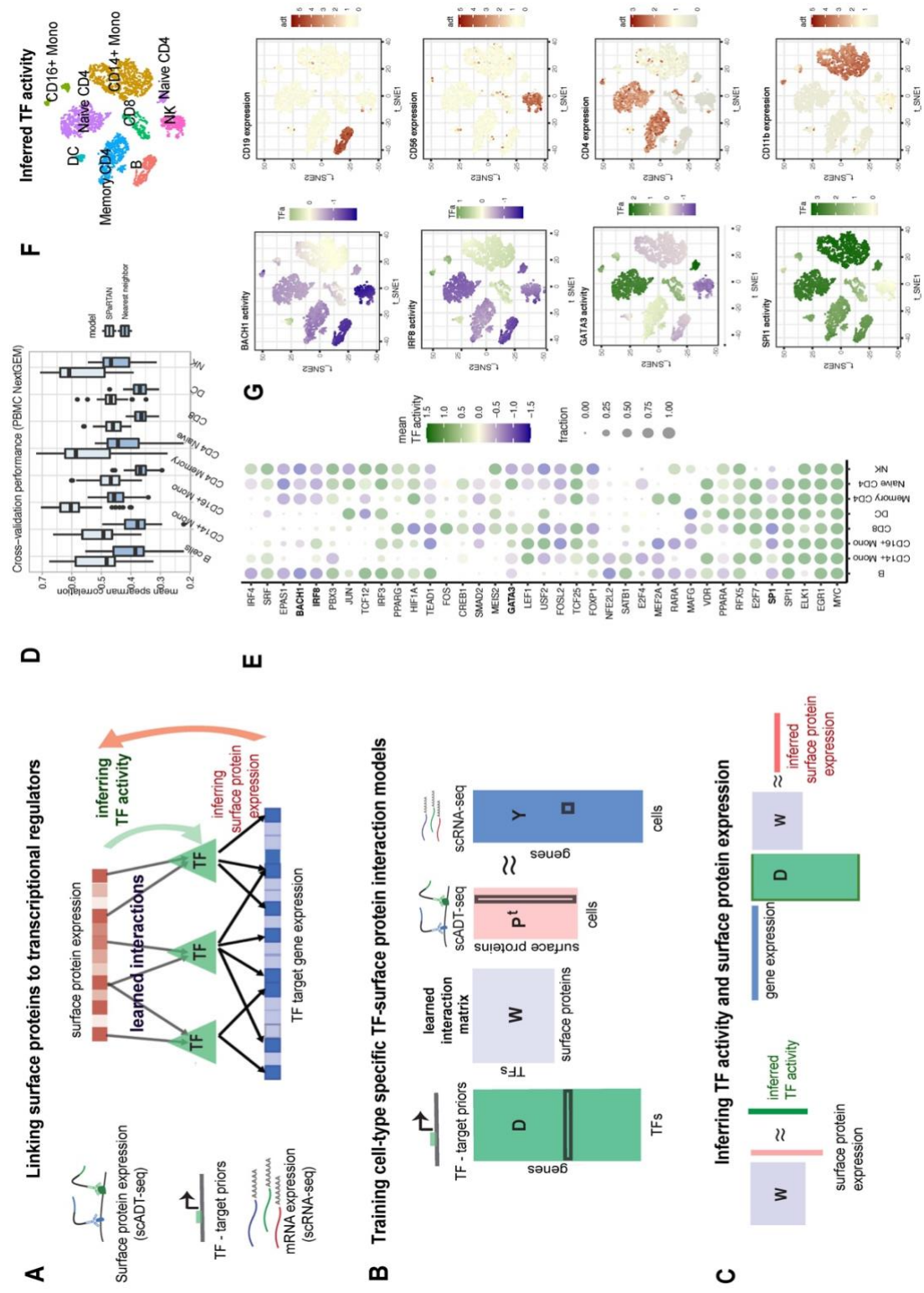
A
Linking surface proteins to transcriptional regulators
Surface protein expression (scADT-seq)
TF - target priors
mRNA expression (scRNA-seq)
surface protein expression
learned interactions
TF
inferring TF activity
inferring surface protein expression
TF target gene expression
B
Training cell-type specific TF-surface protein interaction models
TF - target priors
genes
D
TFs
learned interaction matrix
W
TFs
surface proteins
scADT-seq
surface proteins
cells
scRNA-seq
genes
cells
C
Inferring TF activity and surface protein expression
W
surface protein expression
inferred TF activity
gene expression
D
W
inferred surface protein expression
D
Cross-validation performance (PBMC NextGEM)
mean spearman correlation
model
SPoRTAN
Nearest neighbor
B cells
CD14+ Mono
CD16+ Mono
CD4 Memory
CD4 Naive
CD8
DC
NK
E
IRF4
SRF
EPAS1
BACH1
IRF8
PBX3
JUN
TCF12
IRF3
PPARG
HIF1A
TEAD1
FOS
CREB1
SMAD2
MEIS2
GATA3
LEF1
USF2
FOSL2
TCF25
FOXP1
NFE2L2
SATB1
E2F4
MEF2A
RARA
MAFG
VDR
PPARA
RFX5
E2F7
SP1
SPI1
ELK1
EGR1
MYC
mean TF activity
fraction
F
Inferred TF activity
CD16+ Mono
Naive CD4
CD14+ Mono
DC
CD8
Naive CD4
NK
Memory CD4
B
G
BACH1 activity
IRF8 activity
GATA3 activity
SPI1 activity
CD19 expression
CD56 expression
CD4 expression
CD11b expression
t_SNE1
t_SNE2
TFa
adt

Fig. 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 responses. 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, 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. **(D)** SPaRTAN accurately predicts relative gene expression on held-out PBMC (10X Genomics, NEXTGEM, training dataset) cells for each cell-type. Performance of the SPaRTAN models for each PBMC cell type compared to nearest neighbor methods. 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 10X Genomics (Next GEM) each cell-type. **(E)** 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. **(F)** t-SNE on the inferred TF activity matrix. Cells are colored according to major cell types. **(G)** BACH1, IRF8, GATA3, and SPI1 inferred TF activity and CD19, CD56, CD4 and CD11b protein expression overlay on t-SNE of TF activities.

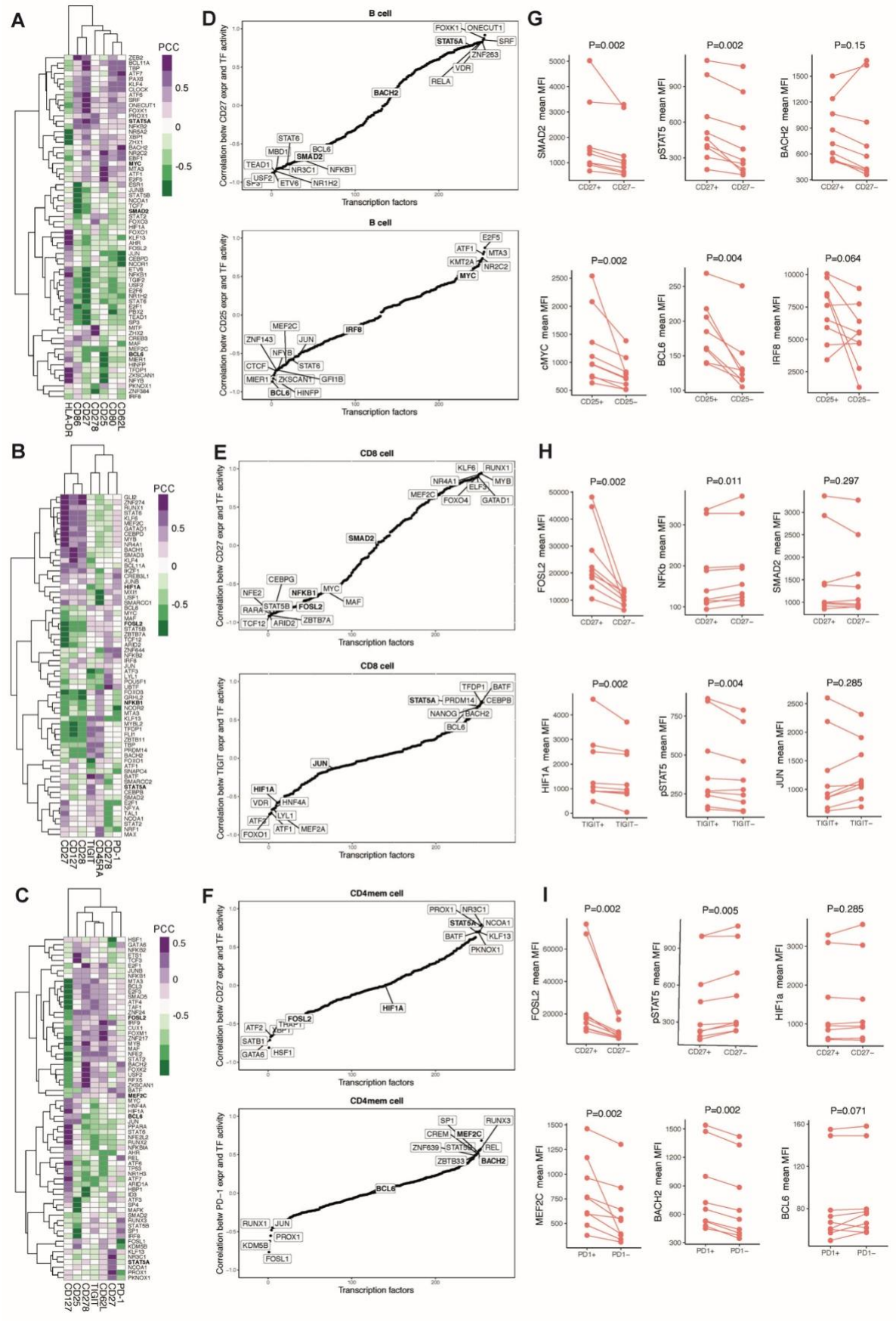
A
B
C
D
E
F
G
H
I
PCC
0.5
0
-0.5
B cell
CD8 cell
CD4mem cell
Correlation betw CD27 expr and TF activity
Correlation betw CD25 expr and TF activity
Correlation betw TIGIT expr and TF activity
Correlation betw PD−1 expr and TF activity
Transcription factors
P=0.002
P=0.15
P=0.004
P=0.064
P=0.011
P=0.297
P=0.285
P=0.005
P=0.071
SMAD2 mean MFI
pSTAT5 mean MFI
BACH2 mean MFI
cMYC mean MFI
BCL6 mean MFI
IRF8 mean MFI
FOSL2 mean MFI
NFKb mean MFI
HIF1A mean MFI
JUN mean MFI
HIF1a mean MFI
MEF2C mean MFI
CD27+
CD27−
CD25+
CD25−
TIGIT+
TIGIT−
PD1+
PD1−

Fig. 2: Analysis of 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)** $CD4^+$ 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.

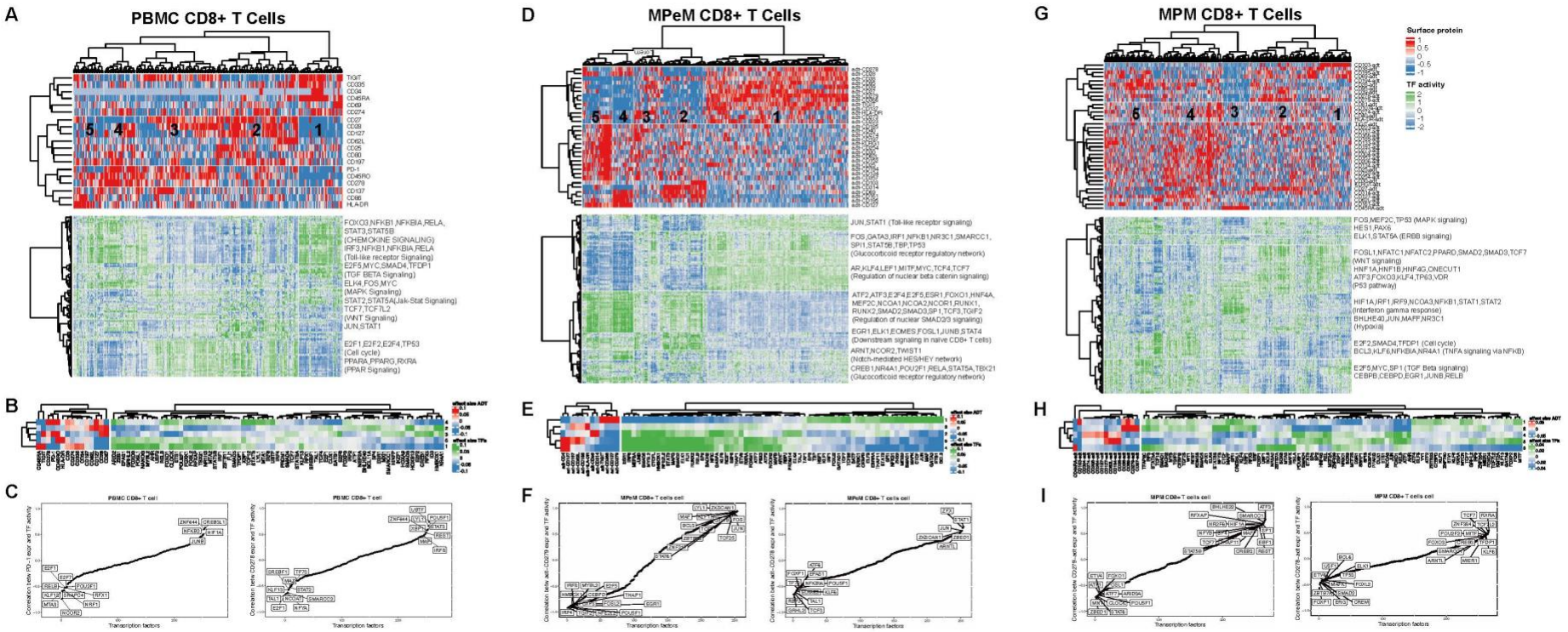
A
PBMC CD8+ T Cells
TIGIT
CD335
CD34
CD45RA
CD69
CD274
CD27
CD38
CD127
CD62L
CD25
CD80
CD197
PD-1
CD45RO
CD278
CD137
CD86
HLA-DR
5 4 3 2 1
FOXO3,NFKB1,NFKBIA,RELA,
STAT3,STAT5B
(CHEMOKINE SIGNALING)
IRF3,NFKB1,NFKBIA,RELA
(Toll-like receptor Signaling)
E2F5,MYC,SMAD4,TFDP1
(TGF BETA Signaling)
ELK4,FOS,MYC
(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
C
PBMC CD8+ T cell
Correlation betw RD - expr and TF activity
Transcription factors
PBMC CD8+ T cell
Correlation betw CD278 expr and TF activity
Transcription factors
D
MPeM CD8+ T Cells
5 4 3 2 1
JUN,STAT1 (Toll-like receptor signaling)
FOS,GATA3,IRF1,NFKB1,NR3C1,SMARCC1,
SPI1,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 receptor regulatory network)
E
F
MPeM CD8+ T cells cell
Correlation betw anti-CD278 expr and TF activity
Transcription factors
MPeM CD8+ T cells cell
Correlation betw anti-CD278 expr and TF activity
Transcription factors
G
MPM CD8+ T Cells
Surface protein
TF activity
5 4 3 2 1
FOS,MEF2C,TP53 (MAPK signaling)
HES1,PAX6
ELK1,STAT5A (ERBB signaling)
FOSL1,NFATC1,NFATC2,PPARD,SMAD2,SMAD3,TCF7
(WNT signaling)
HNF1A,HNF1B,HNF4G,CNECUT1
ATF3,FOXO3,KLF4,TP63,VDR
(P53 pathway)
HIF1A,IRF1,IRF9,NCOA3,NFKB1,STAT1,STAT2
(Interferon gamma response)
BHLHE40,JUN,MAFF,NR3C1
(Hypoxia)
E2F2,SMAD4,TFDP1 (Cell cycle)
BCL3,KLF6,NFKBIA,NR4A1 (TNFA signaling via NFKB)
E2F5,MYC,SP1 (TGF Beta signaling)
CEBPB,CEBPD,EGR1,JUNB,RELB
H
I
MPM CD8+ T cells cell
Correlation betw CD278 - expr and TF activity
Transcription factors
MPM CD8+ T cells cell
Correlation betw CD278 - expr and TF activity
Transcription factors

Fig. 3: 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 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. **(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.