## Supplementary Methods

### Association of sorted proportions with clinical outcome

In order to standardize the estimation of cell proportions across patients, automated gating was performed with R package openCyto v1.22.2 using the singletGate, flowClust.2D, and rectangleGate functions. Fluorescence intensities for cell surface proteins were log-transformed, and batch correction of the CD95 fluorophore was performed by mean-centering. Proportions of naive, stem cell memory, central memory, effector memory, and effector T-cells were calculated by dividing the number of gated cells for each T-cell type by the total number of gated cells.

In order to associate the proportions of each T-cell subset with clinical outcome, we dichotomized the 41 patients from clinical trial NCT02906371, for whom BCA data were uniformly available, into patients with a high or low proportion of each T-cell type, using the median proportion as a cutoff. We performed Cox Proportional Hazards with event-free survival using R package survcomp. Statistical significance was assessed using the log-rank test.

### Bulk RNA-Seq Alignment, Quantification, and Quality Control

Paired-end RNA-Seq reads were aligned with STAR v2.5.3a to the hg38 genome assembly using the Gencode v29 genome annotation. RNA quality control was performed using RSeQC v2.0.0 with custom scripts to evaluate mapping rate, 3’ bias, and percent of reads mapping to exonic regions. BAM files were merged and indexed with SAMtools v1.9. Generation of read count matrices was performed with the Rsubread package v1.34.6. Transcript quantification was

performed with Cufflinks v2.2.1, and conversion to transcripts-per-million (TPM) was performed with a custom R script.

**Gene expression deconvolution using bulk and single-cell references**

In order to assess $CD8^+$ and $CD4^+$ T-cell composition, we performed deconvolution with CIBERSORTx[1] using bulk and single cell references. We ran CIBERSORTx with default parameters and with B-mode batch correction for the LM22[2] and 10X genomics reference peripheral blood mononuclear cell (PBMC) datasets[3], observing strong correlation in $CD4^+$ and $CD8^+$ T-cell proportion estimates. We validated the deconvolution estimates by assessing the proportions of estimated naive, memory, and activated memory T-cells in the LM22 reference compared to the sorted T-cell subsets (Supplemental Figure 3). In order to derive a robust estimate of $CD4^+$ and $CD8^+$ proportions, we computed the mean of the four computational runs as the final consensus estimate.

**Batch effect correction and dimensionality reduction**

TPM values with a pseudocount of 1 were log-transformed, and batch correction for three experimental batches was performed with Limma. Principal component analysis was performed to reduce the dimensionality of the data. Since the first principal component was associated with RNA quality metrics, its effect was linearly regressed from log-transformed expression values. To visualize the low-dimensional data, t-distributed stochastic neighbor embedding (t-SNE) was performed on the first 50 principal components of the RNA-seq expression matrix, with approximately 70% of variance explained.

**Identification of differentially expressed genes**

Read counts were filtered to include only protein-coding genes with a read count of at least 5 in greater than 50% of samples. Patients were dichotomized into those who failed to achieve 6 months of CAR T-cell persistence and those who achieved at least 6 months of CAR T-cell persistence based on the event-free survival criteria. Patients who were observed to have B-cell aplasia that was censored before 6 months were considered to have uncertain categorization and were excluded from differential expression analysis requiring patient dichotomization. Differential expression analysis was performed using a mixed-effects interaction model with Limma-Voom[4,5]. Experimental batch was regressed as a covariate in this model. From this model, treatment contrasts representing the comparison of clinical outcomes for each T-cell type were extracted.

Formally, the design matrix in Limma-voom specified the following model for expected log-counts per million $y_{gi}$ for gene $g$ and RNA-Seq sample $i$:

$$E(y_{gi}) = \mu_{gi} = \sum_{j=1}^{5}\sum_{k=1}^{2} \alpha_{jk} \cdot T_{ij} \cdot P_{ik} + \beta_2 \cdot batch2_i + \beta_3 \cdot batch3_i + \beta_4 \cdot \mathrm{PC1}_i + \beta_5 \cdot \mathrm{CD8/CD4}_i$$

where $\alpha_{jk}$ represent model coefficients, $T_{ij}$ are the indicator variables for T-cell subsets ($T_{ij}$=1 if the $i^{th}$ sample is of T-cell type $j$ for the five T-cell subpopulations $T_N$, $T_{SCM}$, $T_{CM}$, $T_{EM}$, $T_{EFF}$; 0 otherwise) and $P_{ik}$ are indicator variables for clinical CAR T-cell persistence ($P_{i1}$=1 if the $i^{th}$ sample derived from a patient with at least 6 months of CAR T-cell persistence, 0 otherwise; $P_{i2}$=*1* if the $i^{th}$ sample derived from a patient with less than 6 months of CAR T-cell persistence, 0 otherwise). $\beta_2$ and $\beta_3$ represent the model coefficients assigned to the dummy-encoded batch indicator variables, $\beta_4$ represents the model coefficient assigned to the technical effect observed in the first principal component, and $\beta_5$ represents the model coefficient associated with the

estimated CD8/CD4 ratio from CIBSERSORTx. Since five T-cell subsets were sorted for each patient, we addressed within-patient correlations by treating patient ID as a random effect, which constrains the covariance matrix:

$$var(y_{gi}) = \sigma_{gi}^2 R_{gi},$$

where $R_{gi}$ is the block diagonal matrix returned from the duplicateCorrelation function[6].
From this model, contrasts associated with the coefficients $\alpha_{jk}$ were designed to assess for differentially expressed genes. Three major classes of contrasts were performed: comparisons between T-cell subsets (e.g. $T_{EFF}$ vs non-$T_{EFF}$), comparisons of clinical outcome within individual T-cell subsets (e.g. CAR T-cell persistence vs non-persistence in $T_{EFF}$), and comparisons of clinical outcome across individual T-cell subsets (e.g. CAR T-cell persistence vs non-persistence in $T_{EM}$ and $T_{EFF}$). Adjusted P-values were computed via Benjamini-Hochberg correction.

**Pathway enrichment analysis**

Pathway enrichment analysis for differentially expressed genes was performed using Metascape, which uses the hypergeometric test with Benjamini-Hochberg p-value correction and clustering of similar enriched terms[7]. We used the core gene sets in the Metascape database, which includes Gene Ontology processes, KEGG pathways, Reactome gene sets, canonical pathways, and CORUM complexes. Global comparison of T-cell subtypes was performed through an ANOVA F-test in Limma, in which the top 500 genes with FDR < 0.05 were submitted to Metascape. Comparison of clinical T-cell persistence with T-cell subtypes was performed through the mixed-effects Limma model, and genes with FDR < 0.25 were submitted to Metascape.

Top-ranking gene sets were visualized through the single-sample extension of Gene Set Enrichment Analysis (ssGSEA) described by Barbie *et al.*[8] and implemented in R package GSVA v1.32.0[9] with Gene Ontology gene sets from MSigDB[10]. Pairwise comparisons between ssGSEA enrichment scores between T-cell subtypes were performed with Welch's t-test.

**Generation of T-cell specific transcriptional regulatory networks**

Annotation of transcription factors from six databases were downloaded from Lambert et al.[11]. We used our previously described method for identification of key transcription factors combining gene regulatory inference with differential expression analysis[12,13]. Briefly, base transcriptional regulatory networks were generated using top-performing methods in the DREAM5 Network Inference Challenge: CLR, GENIE3, TIGRESS, and Inferelator[14–17]. Protein-coding genes with TPM greater than 1 in at least 20% of samples were included. A consensus transcriptional regulatory network was generated under the Borda Count principle, in which edge weights for each base network were ranked and averaged. The consensus transcriptional regulatory network G is the directed graph produced by the 1,000,000 top-ranking edges.

This process was repeated for six conditions: $G_{T\text{-}cell}$ from the expression profiles of all five T-cell subpopulations; $G_{early}$ generated from the expression profiles of the early-lineage T-cell subtypes ($T_N$, $T_{SCM}$, and $T_{CM}$); $G_{late}$ from the expression profiles of the effector T-cell lineages ($T_{EM}$ and $T_{EFF}$); $G_{TEFF}$ from the expression profiles for $T_{EFF}$; and $G_{regressed}$ from $G_{T\text{-}cell}$ with the T-cell subtype effect regressed using Limma. We generated a public immune T-cell network, $G_{public}$, using expression data compiled by Becht *et al.*[18] from 708 immune cells from 43 microarray datasets.

We defined T-cell specific transcriptional regulatory networks *S* by removing edges that were present in both the T-cell networks and public immune networks:

$S_{T\text{-}cell} = G_{T\text{-}cell} - G_{public}$
$S_{early} = G_{early} - G_{public}$
$S_{late} = G_{late} - G_{public}$
$S_{regressed} = G_{regressed} - G_{public}$
$S_{TEFF} = G_{TEFF} - G_{public}$

In order to validate the network construction, we assessed the connectivity of known key transcription factors involved in T-cell differentiation from Chang et al 2014 in $S_{T\text{-}cell}$. We compared the out-degree of transcription factors of $S_{T\text{-}cell}$ of known T-cell transcription factors compared to the null distribution, and repeated this process with $G_{public}$ as a negative control.

**Identification of putative key regulating transcription factors from bulk RNA-Seq data**

Identification of key transcription factors was performed as previously described[12,13] under the rationale that key regulating transcription factors propagate their regulatory effect to larger fraction of differentially expressed target genes via direct or indirect connections through a network. The weighted distance between two genes *i* and *j* in the T-cell specific transcriptional regulatory network was defined as

$$W(i,j) = 1 - \frac{log_{10}(p_i) + log_{10}(p_j)}{2log_{10}(p_{min})},$$

where $p_i$ and $p_j$ are the differential expression p-values for genes *i* and *j* respectively, and $p_{min}$ is the smallest differential expression p-value among all genes in the network.

For this weighted transcriptional regulatory network, we used Dijkstra's algorithm to calculate the median shortest path between each TF and differentially expressed target genes (FDR < 0.25)

in the network. Normalized regulatory potential was defined as the median shortest path normalized to zero mean and unit variance. Statistical significance was assessed using topology-preserving randomization, in which the median shortest path from each TF to each differentially expressed gene was compared to the empirical null distribution generated by shuffling edge weights in the network. For each network, we generated 1000 shuffled networks in order to compute empirical p-values.

**Analysis of CITE-Seq data**

Demultiplexing and alignment of RNA and antibody-derived tag sequences was performed with cellranger v3.1.0. Low-quality cells were computationally filtered by retaining only those cells with between 300 and 2500 genes in the scRNA-Seq data, greater than 1500 RNA counts, and less than 10% mitochondrial RNA.

CITE-Seq antibody-derived tag counts were normalized with centered log-ratio (CLR) normalization. Single-cell RNA-Seq count matrices for each patient were log-normalized, and the top 2000 variable genes were identified with the variance-stabilizing transformation. Data integration between the six samples was performed on single-cell RNA-Seq matrices using Seurat v3.2.0 using the default anchor-based canonical correlation analysis (CCA) with 30 dimensions, 2000 anchor features, and k.filter = 100. Initial clustering was performed on the integrated single-cell RNA expression matrix with the default Louvain algorithm in Seurat with the top 20 principal components in order to identify CD4+ and CD8+ T-cells, as well as to remove small clusters of CD19-expressing B-cells and low-count cellular debris.

We sought to identify a final set of T-cell clusters by combining unbiased, transcriptome-wide clustering with prior knowledge of cell surface protein and RNA-based T-cell markers. To this end, CD4+ and CD8+ T-cells were separately clustered using the Lovain algorithm with 20 principal components and a relatively high resolution parameter of 0.8, leading to the identification of 21 T-cell clusters. As many of these clusters shared identical T-cell marker profiles, we visualized the mean expression of CITE-Seq antibody and RNA markers, and merged clusters that had similar marker expression and via a between-cluster dendrogram.

Dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) of the top 20 principal components of the integrated scRNA-Seq data with 30 neighbors and 2 components. Single-cell pathway enrichment scores for the gene signatures defined from our bulk RNA-Seq analysis was performed with AUCell v1.6.1[19]. For visualization of CITE-Seq antibody gene expression or AUCell enrichment scores, minimum and maximum values of color gradients were defined using the 5th and 95th percentile values in order to reduce the impact of outlier values.

**Integrative analysis of scATAC-Seq data with CITE-Seq data as a reference**

Demultiplexing of scATAC-Seq reads was performed with cellranger-atac v1.1.0, and alignment and peak calling were performed with BWA and MACS2 using the scATAC-pro pipeline[20] under default parameters. Peaks were merged with the scATAC-pro mergePeaks module, and peak-by-cell matrices with merged peaks were reconstructed with the scATAC-pro reConstMtx module.

Gene-activity matrices for single-cell ATAC-Seq were constructed by summing counts within the gene body and 2kb upstream, as previously described[21]. Integration of scATAC-Seq samples was performed using gene-activity matrices and Seurat v3.2.0 using the default anchor-based CCA method using 30 dimensions, 2000 anchor features, and k.filter=100. In order to integrate CITE-Seq and scATAC-Seq data, transfer anchors were computed using CCA with scRNA-Seq as the reference, and feature imputation was performed using the TransferData function.

The scATAC-Seq data were projected onto scRNA-Seq principal component space using the loadings associated with the top 20 principal components of the scRNA-Seq data. Each scATAC-Seq cell was associated with its 30 nearest scRNA-Seq neighbors in this principal component space using the cosine distance, and T-cell cluster labels were assigned via *k*-nearest neighbors with $k = 30$. For visualization, scATAC-Seq data were projected onto the scRNA-Seq UMAP coordinates using the uwot::umap_transform function.

For TF motif enrichment analysis, we first generated a peak-by-motif matrix using Signac v1.0.0[21] with the JASPAR 2020 database of human transcription factor binding motifs[22], then computed motif activity scores using chromVAR v1.6.0. For UMAP visualizations, the color scales were defined by the 5th and 95th percentile values and centered at zero.

In order to identify high-confidence enhancer-promoter interactions, we used two approaches: chromatin co-accessibility, and a regression-based meta-cell approach integrating the scRNA-Seq and scATAC-Seq data. To assess for chromatin co-accessibility, we ran Cicero v1.3.4.10[23], which outputs a list of connections between chromatin peaks with an associated co-accessibility

score. Putative enhancer-promoter pairs were restricted to those for which one peak overlapped with a gene promoter, and the other did not overlap with a promoter or exonic region. As a second line of evidence supporting enhancer-promoter interactions, we used a regression-based meta-cell based approach developed to integrate information from scRNA-Seq and scATAC-Seq data while addressing sparsity in the count matrices[24]. For each cell in the scRNA-Seq dataset, a scRNA-Seq and scATAC-Seq "meta-cell" was defined by pooling counts for each gene or peak from the 30 nearest neighbors in the principal component space by cosine distance. Meta-cell counts were log-normalized and scaled to zero mean and unit variance. For a gene of interest, we ran a linear regression model using meta-cell gene expression as the dependent variable, and putative enhancer peaks within 200kb of the transcription start site as regressors. Bonferroni adjusted p-values less than 0.05 with a positive coefficient were considered significant. Both Cicero and meta-cell based approaches were run on the entire single-cell T-cell dataset, as well as on a restricted set of $CD8^+$ $T_{EFF}$ in order to characterize the upstream regulation of *TCF7* across T-cell lineages and within the $CD8^+$ $T_{EFF}$ subset.

In order to assess the gene regulation of *TCF7* downstream targets across T-cells and within the CD8+ $T_{EFF}$ subset, we assessed the Cicero chromatin co-accessibility scores for *TCF7* peak-to-promoter pairs and null peak-to-promoter pairs. *TCF7* peak-to-promoter pairs were defined as those Cicero putative enhancer-promoter pairs for which one peak contained the *TCF7* motif, and the other contained the promoter region of a gene within the *TCF7* regulon gene signature defined from the bulk RNA-Seq analysis. Null peak-to-promoter pairs were defined as those Cicero putative enhancer-promoter pairs for which the putative enhancer did not contain the *TCF7* motif, and the promoter peak was not among the *TCF7* regulon genes. Statistical

significance between the Cicero co-accessibility scores of the TCF7 peak-to-promoter pairs and null peak-to-promoter pairs was assessed with the Wilcoxon rank-sum test. This process was performed using Cicero output run on the entire scATAC-Seq dataset, and on strictly the CD8+ $T_{EFF}$ subset.

**3C-qPCR**

3C was performed as described before with minor modifications[25]. Briefly, 0.5~1 million sorted T subpopulations were cross-linked with 2% formaldehyde for 10 min at RT and then quenched by adding 0.125 M glycine. Cell pellets were lysed with 0.5 mL of lysis buffer (10 mM Tris-HCl, 10 mM NaCl, 0.5% NP-40, and 1x protease inhibitor) for 10 min on ice. Nuclei were resuspended with 450 μL of H2O. 60 μL of 10x DpnII buffer and 15 μL of 10% SDS were added and mixed before incubation at 37 °C with shaking at 1,400 rpm for 1 hour. 75 μL of 20% Triton X-100 was added and mixed before incubating at 37 °C with shaking at 1,400 rpm for 1 hour. Three aliquots (250U, 500U, 250U) of DpnII (NEB, Cat #: R0543M) were added to each reaction mix several hours apart with incubation at 37 °C with shaking at 1,400 rpm. DpnII was then heat-inactivated by incubating at 65 °C for 20 min. Samples were subjected to ligation by adding 30 μL of H2O, 70 μL of 10x T4 DNA ligase buffer, and 50 U of T4 DNA ligase (Roche, Cat #: 10799009001). Ligation was done at 16 °C for 8 hours with slow rotation. Samples were reverse cross-linked by incubating with proteinase K (NEB, Cat #: P8107S) at 65 °C overnight and RNase A (Thermo Scientific, EN0531) at 65 °C for 30 min. DNA was purified by phenol-chloroform extraction and ethanol precipitation. 3C controls were amplified from gDNA for *TCF7* locus (See Supplementary Table X for primer information). Amplified 3C control DNA was digested with DpnII (NEB, Cat #: R0543M) for 1 hour at 37 °C. DNA was purified using

MinElute PCR Purification Kit (Qiagen, Cat #: 28004). Digested DNA was then ligated with T4 DNA ligase (Roche, Cat #: 10799009001) 1 hour at 25 °C. Ligated DNA was purified using MinElute PCR Purification Kit (Qiagen, Cat #: 28004). To quantify specific chromatin interactions, relative interaction frequency was calculated using the following formula:

$$2^{-\Delta\Delta Ct} = 2[(Ct^{3C}_{Target} - Ct^{3C}_{Loading}) - (Ct^{control}_{Target} - Ct^{control}_{Loading})]$$

where $Ct^{3C}_{Target}$ and $Ct^{control}_{Target}$ quantify PCR products at the test locus in the 3C and gDNA template, respectively, and $Ct^{3C}_{Loading}$ and $Ct^{control}_{Loading}$ PCR product at internal loading control locus in the 3C and gDNA template, respectively.

**RT-qPCR**

Total RNA was isolated using the RNeasy micro kit (Qiagen, Cat #: 74004) with on-column genomic DNA removal. cDNA was synthesized from total RNA using the iScript cDNA Synthesis Kit (Bio-Rad, Cat #: 1708891) according to the vendor's instruction. qPCR reactions were performed on an Applied Biosystems ViiA 7 real-time PCR system with iTAQ Universal SYBR® Green Supermix (Bio-Rad, Cat #: 1725124). Relative gene expression was calculated using the $2^{-\Delta\Delta Ct}$ method using *ACTB* gene as the reference.

**Flow Analysis**

Sample processing and surface marker antibody staining followed the same protocol as the primary cohort. Surface marker-stained cells were then fixed and permeabilized using True-Nuclear Transcription Factor Kit (Biolegend, Cat # 424401) according to the vendor's instruction. TF (TCF7 and IRF7) antibody cocktail was added to the fixed samples before

incubation at RT for 2 hours. Cells were then washed twice and analyzed on Cytek Aurora flow cytometry.

## References


1. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
2. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
3. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
4. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
5. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
6. Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067–2075 (2005).
7. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
8. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
9. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
10. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
11. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **175**, 598–599 (2018).
12. Gao, L. *et al.* RUNX1 and the endothelial origin of blood. *Exp. Hematol.* **68**, 2–9 (2018).
13. Gao, P. *et al.* Risk variants disrupting enhancers of TH1 and TREG cells in type 1 diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7581–7590 (2019).
14. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **5**, e8 (2007).
15. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* vol. 5 e12776 (2010).
16. Haury, A.-C., Mordelet, F., Vera-Licona, P. & Vert, J.-P. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.* **6**, 145 (2012).
17. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7**, R36 (2006).
18. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
19. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

20. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol.* **21**, 94 (2020).
21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
22. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz1001.
23. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).
24. Zhu, Q. *et al.* Developmental trajectory of prehematopoietic stem cell formation from endothelium. *Blood* **136**, 845–856 (2020).
25. Hagege, H. *et al*. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722 (2007).