# Projecting single-cell transcriptomics data onto a reference T cell atlas to interpret immune responses

Massimo Andreatta[1], Jesus Corria-Osorio[1], Sören Müller[2], Rafael Cubas[3], George Coukos[1] and Santiago J. Carmona[1*]

[1]Ludwig Institute for Cancer Research, Lausanne Branch, and Department of Oncology, CHUV and University of Lausanne, Epalinges, 1066, Switzerland. [2]Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, California, United States. [3]Department of Translational Oncology, Genentech, South San Francisco, California, United States

## Abstract

Single-cell transcriptomics is a transformative technology to explore heterogeneous cell populations such as T cells, one of the most potent weapons against cancer and viral infections. Recent advances in this technology and the computational tools developed in their wake provide unique opportunities to build reference atlases that can be used to systematically compare new single-cell RNA-seq (scRNA-seq) datasets derived from different models or therapeutic conditions. We have developed ProjecTILs (https://github.com/carmonalab/ProjecTILs), a novel computational tool to project new scRNA-seq data into a reference map of T cells, allowing their direct comparison in a stable, annotated system of coordinates. ProjecTILs enables the classification of query cells into curated, discrete states, but also over a continuous space of intermediate states. We illustrate the projection of several datasets from recent publications over two novel cross-study murine T cell reference atlases: the first describing tumor-infiltrating T lymphocytes (TILs), the second characterizing acute and chronic viral infection. ProjecTILs accurately predicted the effects of multiple perturbations, including the ablation of genes controlling T cell differentiation, such as *Tox*, *Ptpn2*, *miR-155* and *Regnase-1*, and identified novel gene programs that were altered in these cells (such as a *Lag3-Klrc1* inhibitory module), revealing mechanisms of action behind these immunotherapeutic targets and opening new opportunities for the identification of novel targets. By comparing multiple samples over the same reference map, and across alternative embeddings, our method allows exploring the effect of cellular perturbations (e.g. as the result of therapy or genetic engineering) in terms of transcriptional states and altered genetic programs.

* e-mail: santiago.carmona@unil.ch

## Introduction

Single-cell RNA sequencing (scRNA-seq) technologies have revolutionized the study of transcriptome-wide gene expression, enabling the measurement of molecular profiles at single-cell resolution[1]. Single-cell technologies offer unique opportunities to understand complex biological systems – such as the study of cell differentiation or tumor heterogeneity, or to assess the effect of therapeutic interventions on the phenotype of cell populations[2]. One of the most ambitious projects is perhaps the construction of a Human Cell Atlas, an international effort aimed at characterizing the molecular profiles of cells in all human tissues[3].

For the study of immune responses and design of cancer immunotherapies, single-cell profiling of T cell transcriptomes has become a powerful tool to describe the diversity of T cell states, and their relation to disease progression and response to therapy[2,4]. While the presence of tumor-infiltrating T lymphocytes (TILs) in cancer lesions has been broadly associated with improved prognosis and response to immune checkpoint blockade[5,6], scRNA-seq has uncovered a great diversity within TILs, revealing that distinct TIL states contribute differently to tumor control and response to immunotherapies[7–9]. Crucially, antigen persistence in cancer and chronic infections profoundly alters CD8 T cell differentiation and function, leading antigen-specific cells into a collection of states referred to as "exhaustion"[10]. Recent data has shown that exhausted CD8 T cells are phenotypically and functionally heterogeneous[11–17]. In the tumor microenvironment (TME), at least two epigenetically distinct states of exhausted CD8 T cells exist: the PD-1+ TIM-3- TCF1+ "progenitor exhausted" memory-like state with self-renewal capacity, and the PD-1+ TIM-3+ TCF1- "terminally exhausted" effector state with little proliferative potential and higher cytolytic activity[15,16]. CD4 TILs are also highly heterogeneous and exert multiple and opposing roles in tumor control. For instance, Th1 cells produce IFN-γ and exert prominent antitumor activity[18]; conversely, T regulatory cells (Treg) inhibit antitumor responses by different mechanisms including CD8 T cell suppression, and high Treg infiltration in the TME is often associated with poor prognosis and survival[5]. Such diversity can now be studied at single-cell resolution thanks to scRNA-seq technologies and offers multiple entry points for therapeutic intervention aimed at boosting antitumor responses. However, comprehensive cross-study reference maps of T cell types/states, encompassing their transcriptional diversity, are still lacking.

A major challenge towards the construction of single-cell atlases is the integration of gene expression datasets produced from different samples, from different tissues or using different technologies. Several computational methods have been developed to correct technical biases introduced by handling experiments in "batches", and to align datasets over their biological similarities. For example, Harmony[19] uses an iterative approach to cluster cells into cell types and calculate correction factors for each cell with the goal to minimize batch effects. scAlign[20] relies on deep neural networks to calculate a bidirectional mapping between pairs of datasets, in this way allowing the comparison of datasets over shared coordinate spaces. Two recently published tools for batch-correction, Seurat3[21] and LIGER[22], allow integration across different modalities, for example between single-cell transcription

levels and epigenetic profiles, as measured by chromatin accessibility. We recently proposed STACAS[23], a tool for data integration based on Seurat3 and specifically designed for the challenges of integrating heterogeneous datasets characterized by limited overlap of cell subtypes. This is particularly relevant towards the construction of a T cell atlas, where differences between datasets are not merely the result of technical variation of handling samples in different batches, but rather due to subpopulations of highly variable frequency, and in many cases cell subtypes that are entirely missing from one or more samples, as a result of study design or biological context. While whole-organism single-cell atlases are very powerful to describe the global properties of cell populations, only by constructing specialized atlases for individual cell types can one achieve the level of resolution required to discriminate all previously identified transcriptional states that can be assumed by each cell type.

A second outstanding challenge in single-cell data science is the mapping of single cells to a reference atlas[24]. The ability to embed new data points into a stable reference map has the advantage of robust, reproducible interpretation of new experiments in the context of curated and annotated cell states. In addition, it allows comparing different conditions (e.g. pre- vs. post-treatment, or mutant vs. wild-type) over a unified transcriptomic reference landscape. Because new cells are assigned coordinates in the unified space of the reference atlas, projection of a query dataset allows interpreting the transcriptional state of new cells in a continuous rather discrete space, thereby capturing intermediate as well as transient cellular states. In the absence of reliable reference maps, and computational tools to project new data into a reference map, researchers must rely on unsupervised, manual annotation of their data, a time-consuming and, to a certain degree, subjective process. In this work, we describe ProjecTILs, a computational framework for the projection and embedding of scRNA-seq data into a reference map of T cell states, enabling: *i)* classification of T cells into discrete, curated states, as well as *ii)* in a continuum of intermediate states along the main axes of variation of the reference map; *iii)* automated analysis of T cell perturbations; *iv)* identification of genetic programs significantly altered in a query dataset, within individual cell subtypes, compared to the baseline reference map or to a control dataset. We demonstrate the usefulness of the ProjecTILs algorithm to interpret the effects of multiple T cell perturbations in the context of cancer and infection by a systematic re-analysis of published single-cell data.

# Results

## A cross-study reference atlas of tumor-infiltrating T cell states

With the goal to construct a comprehensive reference atlas of T cell states in murine tumors, we collected publicly available scRNA-seq data from 17 melanoma and colon adenocarcinoma tumors from five different studies (see Methods). Additionally, we generated new scRNA-seq data from 4 tumor-draining lymph node samples (MC38_dLN dataset, see Methods). After data quality checks and filtering pure $\alpha\beta$ T cells, our database comprised expression profiles of 16,803 individual cells from 21 samples.

Substantial batch effects are typically observed between single-cell experiments performed on different days, in different labs, or with different single-cell technologies. Without batch effects correction, cells tend to cluster together by study rather than by cell type (Figure 1A). We applied the STACAS algorithm[23] to integrate the datasets over shared cell subtypes and combine them into a unified map (Figure 1B). Unsupervised clustering and gene enrichment analysis, supported by T cell supervised classification by TILPRED[25] (Figure 1C), allowed annotating areas of the reference map into "functional clusters" (Figure 1D), characterized by known gene expression signatures of specific T cell subtypes. In particular, we observed a distinct separation between CD4 and CD8 T cells, which in turn could be subdivided into subgroups. In particular, we identified a cluster of Naive-like CD8 T cells and a smaller one of Naive-like CD4 T cells, co-expressing *Tcf7* and *Ccr7* while lacking any activation features; a cluster of Effector Memory-like CD8 T cells, co-expressing *Tcf7* and granzymes (most prominently *Gzmk)*, with low to intermediate expression of *Pdcd1*; an "Early activation" state of CD8 T cells, with an intermediate profile between the Naive-like and the EM-like CD8 types; a CD8 Terminally-Exhausted (Tex) effector cluster, characterized by high expression of granzymes, multiple inhibitory receptors (*Pdcd1*, *Ctla4*, *Lag3*, *Tigit*, *Havcr2*/TIM-3, etc.) and *Tox*[25,26]; a CD8 Precursor-Exhausted (Tpex) cluster, with co-expression of *Tcf7*, *Pdcd1*, *Ctla4*, *Tox* but not *Havcr2* or granzymes [25,27,28]; a cluster of CD4 Th1 cells, expressing IFN-gamma receptor 1 (*Ifngr1*) and *Fasl*[29]; a CD4 follicular-helper (Tfh) population[29,30], with a pronounced expression level of *Cxcr5*, *Tox* and *Slamf6*; and a cluster of regulatory T cells (Treg), identified by *Foxp3* (Figure 1D-F). As expected, while TIL samples were enriched in Tex, Tpex and Tregs, tumor-draining lymph nodes were enriched in naïve-like and follicular helper cells (Figure 1B and Suppl. Figures S1-S2). The reference TIL atlas can be explored online at http://TILatlas.unil.ch

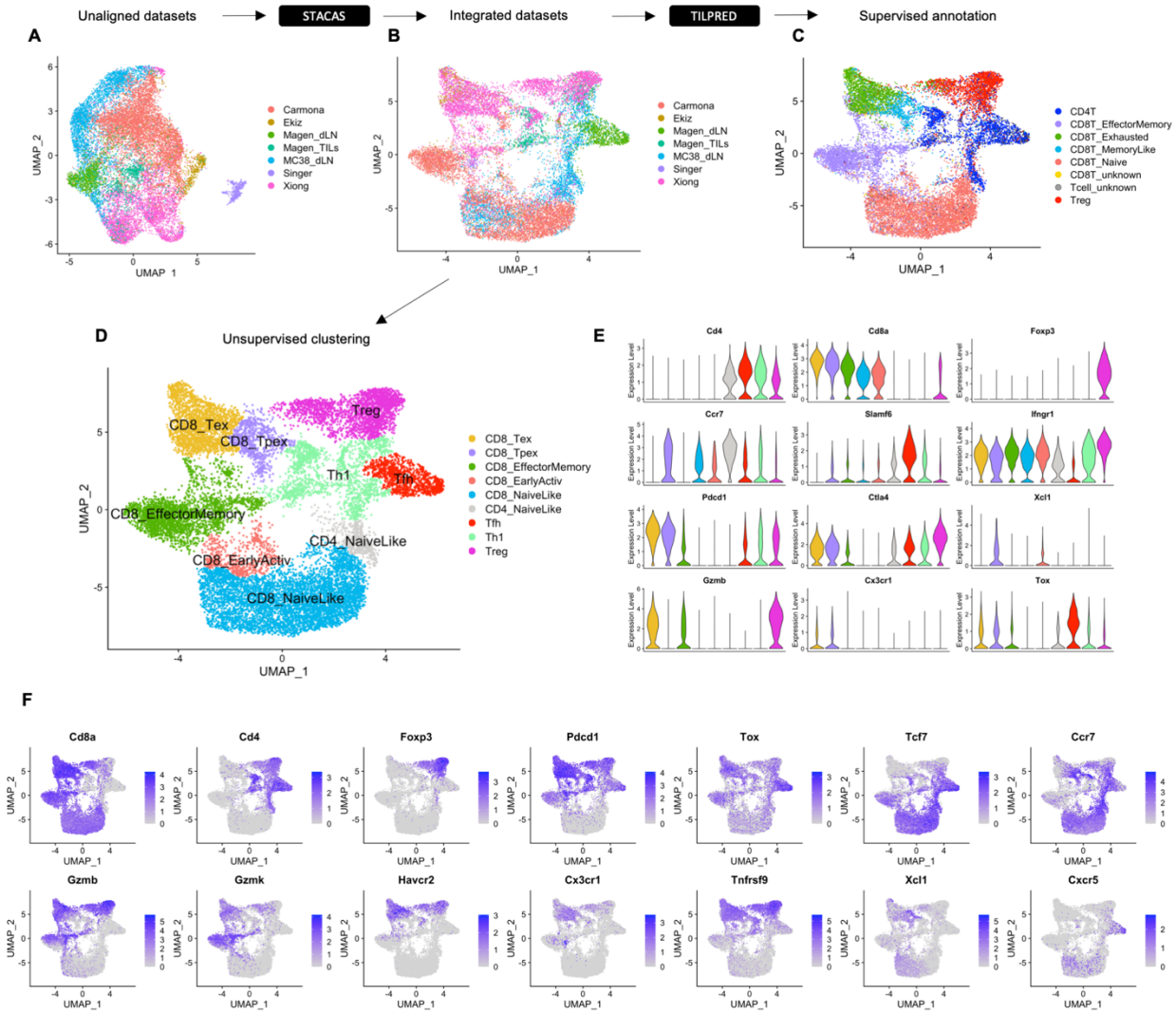



**Figure 1. Building a reference map of TIL transcriptomic states. A)** Uniform Manifold Approximation and Projection (UMAP) reduced dimension plots of single-cell transcriptomic profiles from different studies, before batch effect correction (*i.e.* unaligned datasets); **B)** Same plot for integrated datasets after STACAS alignment: successful dataset integration mitigates batch effects while preserving biological differences between T cell subtypes; **C)** Supervised T cell state classification by TILPRED shows that, after alignment, cells cluster mainly by cell state/subtype, regardless of the dataset of origin; **D)** Unsupervised clusters were annotated as nine functional states based on TILPRED prediction, as well as by **E)** average expression of marker genes in each cluster and by (**F**) single-cell expression of key marker genes over the UMAP representation of the map. An interactive reference TIL atlas can be explored online at http://TILatlas.unil.ch

## ProjecTILs enables accurate projection of scRNA-seq data onto the reference TIL atlas

In order to utilize the constructed reference map not only as a characterization of T cell states in the current datasets, but to also enable interpretation of new datasets, we developed ProjecTILs, a novel method to project new scRNA-seq data onto a reference atlas. The essential input to ProjecTILs is the expression matrix of the query dataset, e.g. in UMI counts or TPMs, where rows represent the genes and columns represent the individual cells. The pre-processing steps (Figure 2A) normalize the data using a log-transformation (if provided non-normalized) and filter out non-T cells (see Methods). In order to reduce batch-effects between the query and the reference map, the STACAS-Seurat integration procedure[23] is used to align the query to the reference, and in this way correct the expression matrix of the query dataset (Figure 2B, see Methods). The corrected query matrix, now in the same scale as the reference map, can then be projected onto reduced dimensionality representations (e.g. PCA, UMAP) of the reference, effectively bringing them into the same reference space. To accomplish this, the algorithm computes the PCA rotation matrix of the reference map, which contains the coefficients to linearly transform gene expression into PCA loadings; the same PCA rotation matrix is then also applied to the query set (Figure 2C). Likewise, the UMAP transformation (allowing the computation of UMAP coordinates from PCA loadings) is applied to the query set to project it into the original UMAP embedding of the reference map. Note that while the expression counts of the query set and their embedding into reduced spaces are modified by the alignment procedure, those of the reference map are not, and its low-dimensional visualizations remain unaltered; different query sets, or experiments containing different conditions, can, therefore, be compared over the same reference map. After projection, a nearest-neighbor classifier predicts the state of each query cell by a majority vote of its annotated nearest neighbors (either in PCA or UMAP space) in the reference map. Benchmarking ProjecTILs by cross-validation experiments showed a high accuracy (>90%) both for classification and projection tasks (Figure 2G, H, Suppl. Figures S1 to S4, and Methods).

As a typical application of ProjecTILs to analyse a query scRNA-seq dataset, we projected onto the reference murine TIL map a dataset of tumor-specific CD8 T cells isolated by tetramer staining from B16-OVA melanoma tumors at day 20 after implantation, from the study by Miller *et al.*[16]. ProjecTILs assigned the great majority (90.2%) of these cells to the Tex state, while small fractions were assigned to the Tpex (4.4%) and EM-like (3.9%) compartments (Figure 2D). Inspecting the average expression for a panel of key markers reveals that the expression profiles of projected cells match well with those of the reference map in different cell states, with a pronounced overexpression of *Gzmb* and the proliferation marker *Mki67* (Figure 2E), as expected by tetramer-induced activation[31,32]. ProjecTILs allows visualizing an additional dimension on the z-axis together with the 2D UMAP representation. In this case, we chose to plot the cell cycling score calculated by TILPRED, which reveals a striking proliferative signal for the cells in the query dataset (Figure 2F). Taken together, these results are well compatible with a population of tetramer+, tumor-specific cells that are highly activated, rapidly dividing and converging to an exhausted phenotype as a result of continued interaction with the tumor.

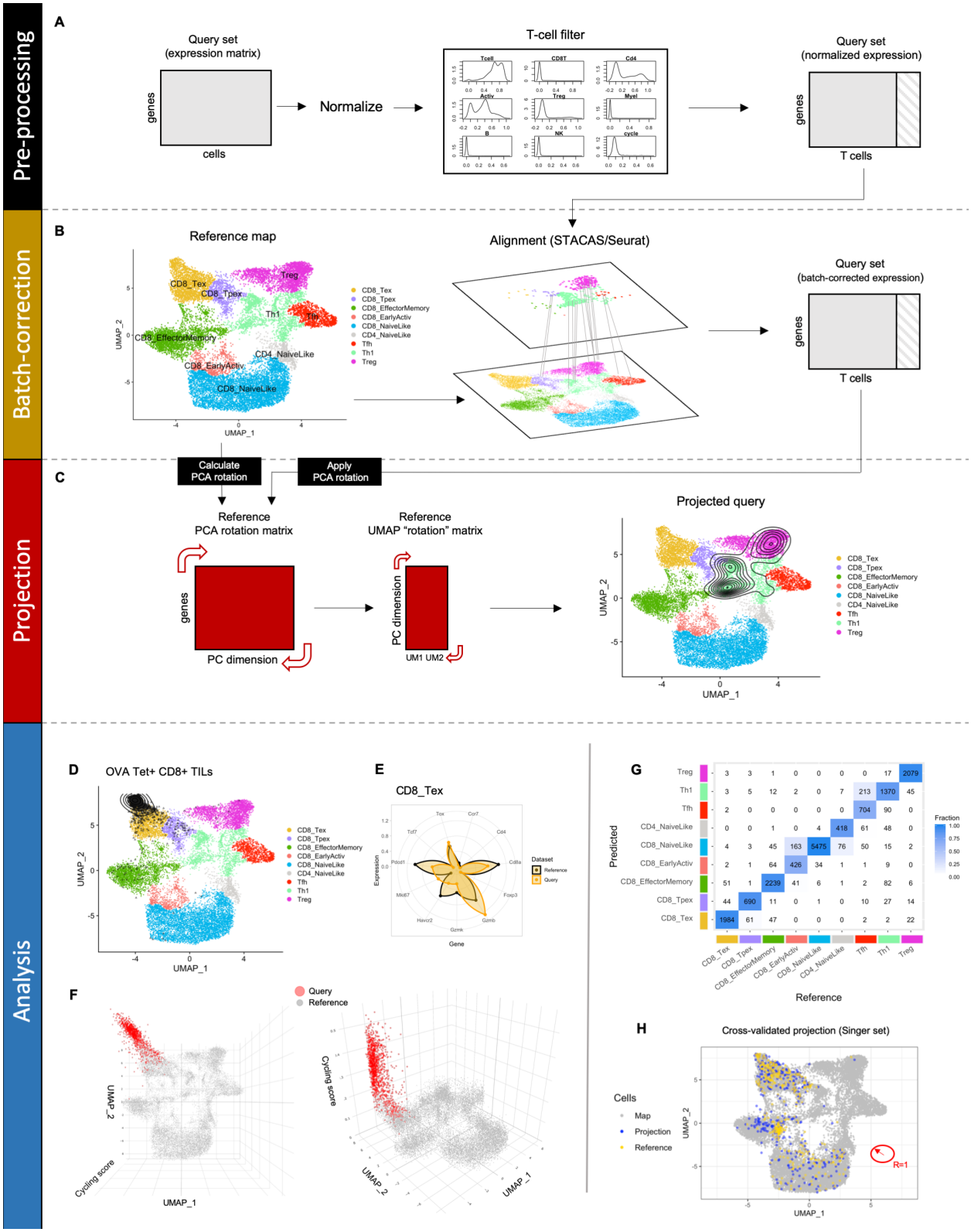



**Figure 2: The ProjecTILs analysis workflow. A)** The essential input to ProjecTILs is a query dataset in the form of a gene expression matrix. Pre-processing steps include data normalization and filtering of non-pure T cells. **B)** The normalized, filtered gene expression matrix is aligned to the reference map using STACAS, to bring the query data into the same scale as the

reference map. **C)** The PCA rotation matrix and UMAP transformation calculated on the reference map are applied to the query set, effectively embedding it into the same space of the reference map, and allowing their direct comparison and joint visualization. **D-F)** Projection example of tumor-specific tetramer+ CD8 TIL single-cell data from Miller *et al.*[16]: **D)** Predicted coordinates in UMAP space as density contours, **E)** gene expression signature for predicted CD8 terminally exhausted (Tex) cells, and **F)** cell cycling score as an additional dimension over the UMAP plot. **G)** Confusion matrix for ProjecTILs state classification in the cross-validation benchmark, comparing reference (x-axis) vs predicted (y-axis) states (the diagonal contains the correctly classified cells, accounting for 91.6%), and **H)** Reference and predicted UMAP coordinates for a representative dataset; the red circle represent the size of a cell neighborhood in UMAP space (of radius 1) to assess projection performance.

## ProjecTILs reveals altered states and gene programs following T cell perturbations

To illustrate the utility of ProjecTILs to investigate the effect of T cell perturbations, such as therapeutic interventions or genetic modifications, we applied the algorithm to the data from Ekiz *et al.*[33], who sequenced total immune (CD45+) single cells from B16 melanoma tumors growing in microRNA 155 (*miR-155*) T cell conditional knock-out (KO) mice, as well as in wild type (WT) mice. It has been previously established that CD8 T cells depend on *miR-155* to acquire anti-tumoral or anti-viral effector functions[34,35]. To understand the impact of *miR-155* deficiency on the TIL landscape, we submitted to ProjecTILs the complete scRNA-seq dataset – which includes B cells and dendritic cells, among others; non-T cells are automatically detected and removed by ProjecTILs before downstream analyses. The population of T cells detected for the WT dataset (16% of the CD45+ total) was predominantly found, as expected, in the Tex compartment. Instead, TILs from the KO mice (8% of the CD45+ total) were mostly projected in the naive-like compartment (Figure 3A), consistent with the critical role of *miR-155* for T cell activation and differentiation. The frequency of cells projected into other T cell types/states was mostly unaltered. Consistently with the predicted change in the dominant T cell phenotype from exhausted to naive-like upon *miR-155* KO, the expression of activation markers (e.g. *Tnfrsf9* and *Ifng*) was reduced in the KO mice, while memory/naive markers such as *Tcf7* and *Ccr7* were overexpressed (Figure 3B). The KO TILs also scored lower in terms of the cycling signature (Figure 3C), further indicating a resting phenotype that does not significantly react to the tumor. In brief, by comparing the changing landscape and cell states distribution of the KO compared to the WT TILs, ProjecTILs provided a straightforward interpretation of the effect of the genetic alteration, all in the context of annotated, reference T cell states and genetic programs. We will discuss additional case studies where ProjecTILs can be employed to characterize novel cell states in later sections.

While low-dimensional representations such as the UMAP are tremendously useful to summarize the most prominent transcriptional features that discriminate distinct T cell subsets, they are often not sufficient to fully capture the heterogeneity of cell transcriptomes. With the goal to provide more resolution to the analysis of T cell states and in particular, to identify gene programs that can be shared by multiple cell subtypes, we decomposed the reference atlas into 50 dimensions using Independent Component Analysis (ICA) (see Methods). Because ICA finds a representation of the data where the dimensions share minimal mutual information, it can be useful to

separate different gene modules and transcriptional programs within complex gene expression datasets[36] and provides a complementary description to the UMAP representation – which in turn is built over a PCA reduction of the cell/gene space. In order to interpret the biological relevance of the ICA components in the reference TIL atlas, we investigated their correlation with annotated molecular signatures from the mSigDB database[37] and observed several modules associated with key cellular pathways (Suppl. Figure S6). For example, the component ICA 33 was driven by several genes associated with hypoxia (e.g. *Tpi1*, *Pkg1*, *Ldha*, *Slc2a1*); ICA 40 was rich in E2F targets (*Mcm2* to *Mcm7*, *Dut*, *Cdc6*), indicating a module of key regulators of cell cycle progression; ICA 26 contained multiple genes involved in cytotoxicity, such as *Prf1* (encoding perforin 1) and several granzymes (*Gzme*, *Gzmc*, *Gzmb*); and ICA 37 was strongly associated with response to interferons (e.g. *Ifit1*, *Ifit3*, *Rsad3*) (Suppl. Figure S7). Critically, some ICA components appeared to affect mainly specific T cell subtypes or regions in the context of our reference map (*e.g* ICA 43 and 45), while others captured features of multiple subtypes/regions (*e.g.* ICA 33 and 37, Suppl. Figure S8).

Importantly, we can compare a projected query dataset with the reference map – or two query conditions between themselves – in their ICA representations, and identify components in which the query dataset deviates from the reference. ICA dimensions where the two sets deviate significantly suggest gene modules that are up- or down-regulated in the query set, and may aid the biological interpretation of phenotypic and functional experimental observations. As an illustrative example of this approach, we re-analyzed the scRNA-seq from Wei *et al*.[38]. In this study, the authors show that KO of *Zc3h12a* (which encodes Regnase-1) in CD8 T-cells improved the therapeutic efficacy of adoptively transferred, tumor-specific cells in mouse models of melanoma and leukemia. Projection of the Regnase-1-null dataset and of a WT counterpart into our TIL reference atlas showed that cells from the two conditions occupied similar CD8 regions of the map (Figure 3D). However, Regnase-1-null T cells displayed a 3-fold enrichment in Tpex and a 1.8-fold enrichment in Tex compared to WT T cells, with a corresponding reduction in the Naive-like and EM compartments (Figure 3E). Interestingly, despite an overall shift towards more differentiated states, compatible with tumor control, the acquired states showed overall increased expression of *Tcf7*, *Slamf6* and *Pdcd1* (which might explain the improved *in vivo* persistence of these cells), as well as lower expression of *Klrc1* (encoding the checkpoint molecule NKG2A, which might explain the enhanced cytotoxic functions) (Figure 3F). Interestingly, discriminant analysis between the ICA dimensions of the two conditions revealed that the most significant deviation was in component ICA 25 (KS test statistic = 0.612, p-val=0), a component driven by several checkpoint molecules such as *Lag3*, *Pdcd1* and *Klrc1* (Figure 3G-H). Visualizing the ICA 25 coordinates of the Regnase-1 KO cells on the z-axis of the UMAP plot, we clearly observe that this component is highly down-regulated compared both to the WT data and the reference map (Figure 3I). While analyzing the expression levels of key marker genes can suggest global patterns of alteration due to the KO (Figure 3F), only by comparing gene programs in each cell subtype individually across conditions (and their relative change in frequency across conditions) can one pinpoint the molecular programs affected by the genetic perturbation while avoiding confounding effects due to global shifts in relative proportions of T cell subtypes. Importantly, a

similar distribution of predicted cell states and significant down-regulation of ICA 25 could be detected even in the absence of a control sample, by direct comparison of the Regnase-1 KO cells to the reference map (Suppl. Figure S9).

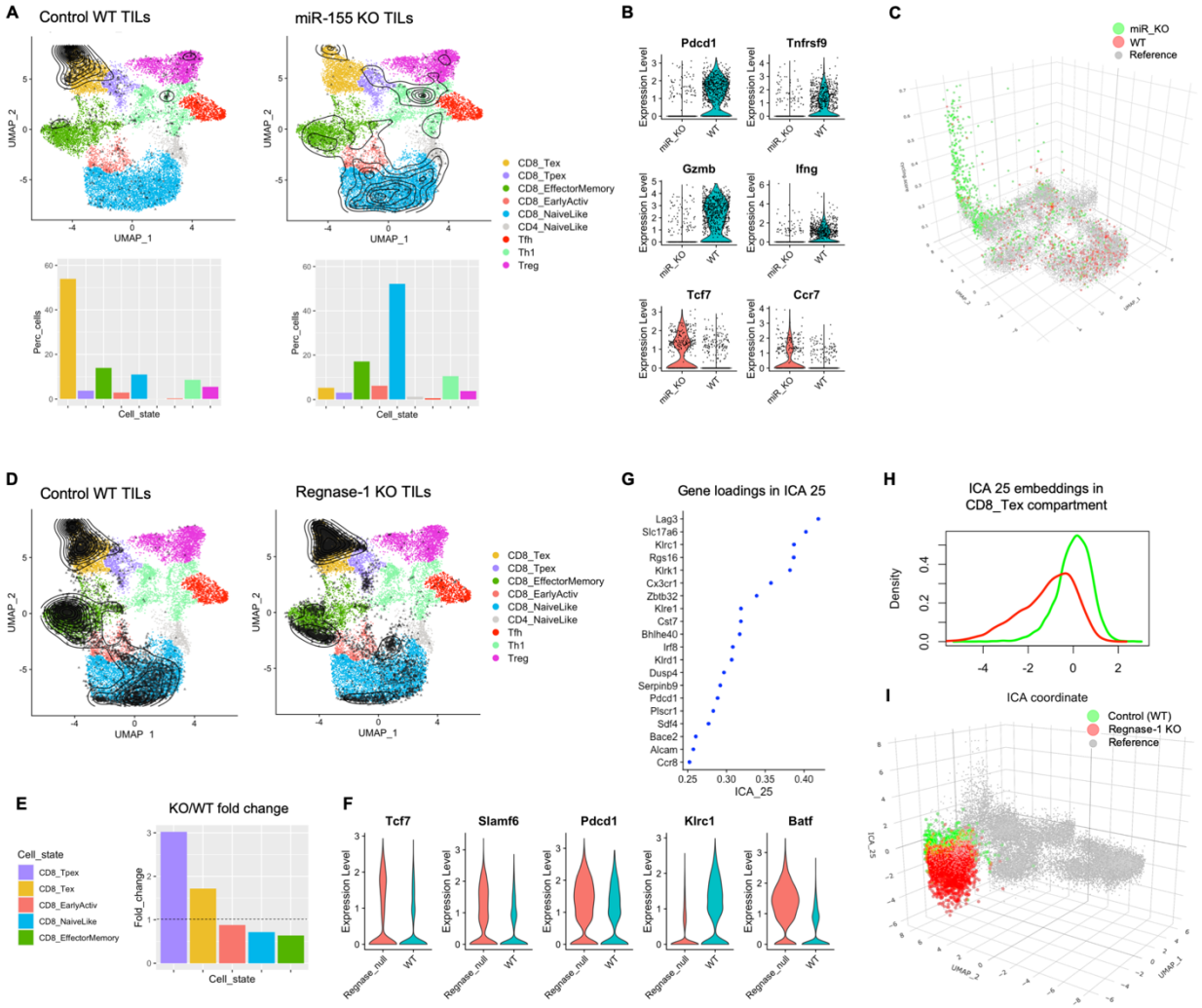



**Figure 3: Investigating the effect of genetic perturbations on T-cell transcriptomes and phenotypes. A-C)** ProjecTILs analysis of the tumor CD45+ scRNA-seq data by Ekiz *et al.*[33]: **A)** WT and *miR-155* KO TILs projected on the reference atlas (black points and density contours) and barplots depicting percentage of cells projected in each T cell state for the two conditions; **B)** Violin plots showing expression of activation and cytotoxicity (*Pdcd1*, *Tnfsf9*/4-1BB, *Gzmb*, *Ifng*) and naïve/memory (*Tcf7*, *Ccr7*) markers; **C)** Cell cycling score represented on the z-axis of the UMAP for the reference map of WT cells and *miR-155* KO cells. **D-I)** ProjecTILs analysis of the scRNA-seq data by Wei *et al.*[38]: **D)** Cells projection on reference TIL atlas (similar to A); **E)** Fold-change in the *Regnase-1* KO compared to WT for each TIL state containing at least 50 cells; **F)** Global expression level for selected genes in the *Regnase-1* KO versus the WT; **G)** Top 20 driver genes in terms of gene loadings for the transcriptional program ICA 25, the most discriminant dimension between the WT and the *Regnase-1* KO; **H)** Distribution of ICA 25 coordinates for cells in the WT (green) versus *Regnase-1* KO (red) samples, as can be also visualized (**I**) by plotting these values on the z-axis of the UMAP plot.

## Custom reference atlases: the transcriptomic landscape of T cell differentiation in acute and chronic infection

While ProjecTILs was originally conceived to study the diversity of TILs, it can be readily applied to any other biological context for which a reference atlas can be constructed. A prominent example is the lymphocytic choriomeningitis virus (LCMV) infection model, one of the best-studied models of acute and chronic viral infection. In order to construct a reference atlas for LCMV infection, we collected scRNA-seq data of virus-specific CD8 T (P14) cells in five batches from three different studies[39–41], consisting of single-cell gene expression measurements at different time points for acute and chronic LCMV infection (Figure 4A). Alignment by STACAS (Figure 4B) was followed by unsupervised clustering (Figure 4C, see Methods). By inspecting the gradients of gene expression across the UMAP representation of the atlas (Figure 4D), as well as the average expression of a panel of marker genes in the different unsupervised clusters (Figure 4E), we annotated seven functional clusters: Effector Early, Effector Intermediate, Effector Cycling, Memory Precursor, Short-Lived Effector Cells (SLEC), Precursor Exhausted (Tpex), and Exhausted (Tex) (Figure 4C). Cells from early acute infection (day 4.5) were mostly located in the Early, Intermediate and Cycling Effector areas, as well as in the Memory Precursor state; however, at a later time point (day 7.5) their distribution of cell states shifted towards the SLEC subtype (Figure 4F). Similarly, early chronic infection (day 4.5) was characterized by Effector T cell types, nearly indistinguishable from the acute cells at the same time point; but as the infection progressed (days 7.5 and 30) their state distribution diverged from the acute infection towards Tpex and Tex states (Figure 4G) with a non-persistent wave of SLEC-like cells at an intermediate time point (day 7.5).

With this reference LCMV map in hand, we proceeded to project new datasets onto the atlas to study the effect of genetic alterations on CD8 T cells during viral infection. The phosphatase PTPN2 has been recently proposed as an attractive immunotherapeutic target to enhance T cell cytotoxicity in chronic infection and cancer[40,42]. Automated ProjecTILs analysis of co-transferred *Ptpn2*-KO and control P14 cells at day 30 after LCMV infection and CD4 depletion[40] revealed that the large majority of cells, both *Ptpn2*-KO and WT, were projected either in the Tpex or Tex clusters (Figure 4H). Importantly, while the Tex compartment constituted only about 20% of the total number of cells in the WT sample, it amounted to over 50% of *Ptpn2*-KO cells (Figure 4I). Moreover, analysis of the average expression of key marker genes confirmed a good agreement between Tpex expression profiles (for control, KO and reference state), as well as overexpression of *Pdcd1* and *Tox* in the *Ptpn2*-null cells (Figure 4J). Therefore, ProjecTILs automated analysis supported the original observations that *Ptpn2* deletion promotes differentiation of Tpe into Tex, which translates into a higher number of effector cells, improving – at least transiently – viral and tumor control[40].

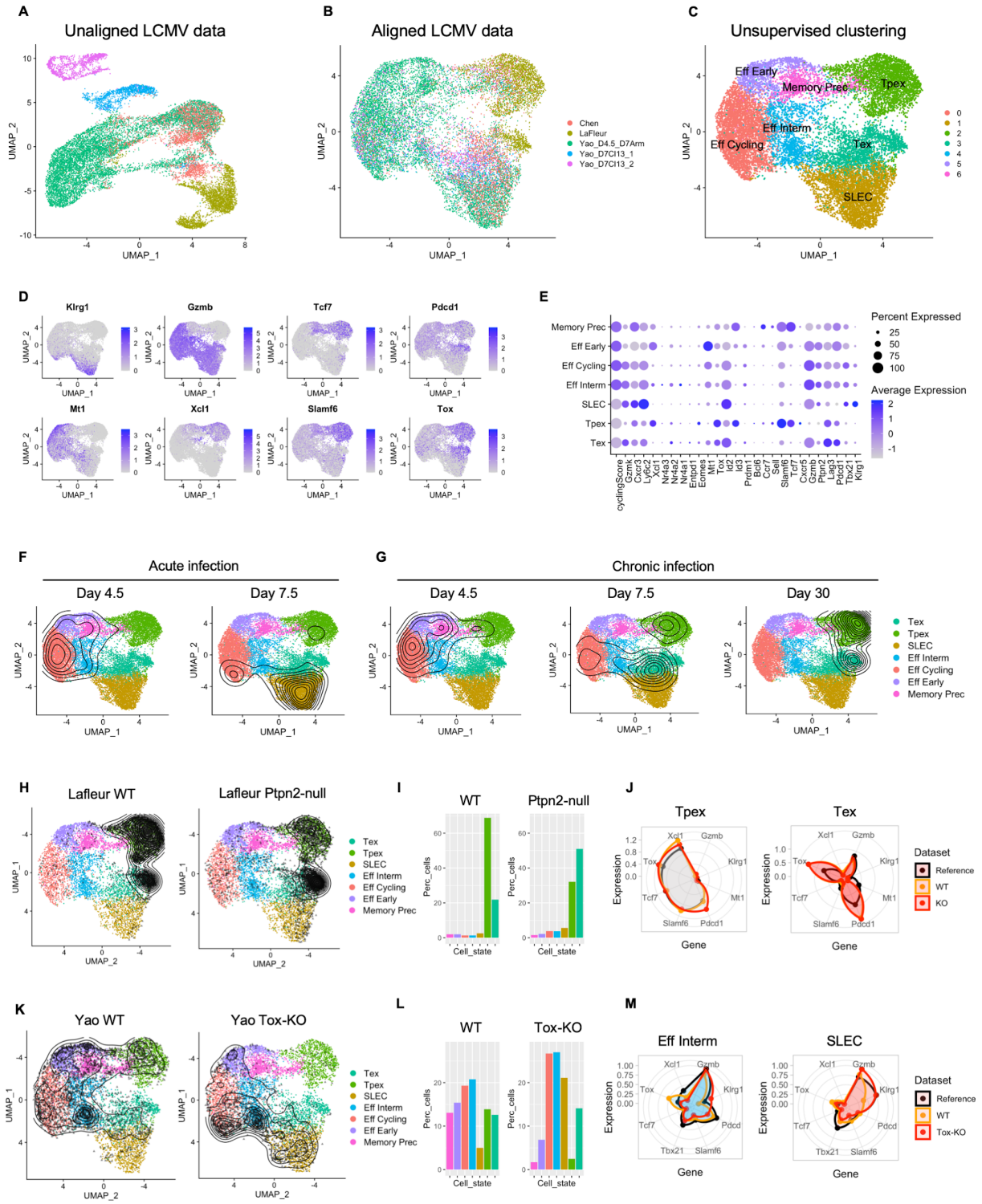



**Figure 4: Beyond TILs: Reference atlas of CD8 T cells during LCMV viral infection. A)** Unaligned datasets of virus-specific (P14) CD8 T cell during infection show pronounced batch effect, which **B)** can be mitigated by STACAS alignment. **C)** Unsupervised clusters were annotated to seven functional clusters by examining **D)** the gradient of expression and **E)** the average expression of marker genes by cluster (*i.e.* Memory Precursors; "Early", "Cycling", "Intermediate", and short-lived ("SLEC") effectors;

Precursor Exhausted (Tpe) and Terminal exhausted (Te) CD8 T cells). **F)** Density of cells across the map at two time points in acute infection and **G)** at three different time points in chronic infection. **H-J)** Analysis of *Ptpn2* KO versus control (WT) using the data by Lafleur *et al.*[40]: **H)** ProjecTILs projection of WT and *Ptpn2* KO cells onto the infection reference map; **I)** Predicted percentage of cells for each T cell state; **J)** Normalized average expression for selected markers in the reference map, in WT and *Ptpn2* KO cells. **K-M)** Analysis of *Tox* KO versus control (WT) using the data by Yao *et al.*[39]: **K)** Projection in UMAP space by ProjecTILs for the WT and *Tox* KO samples; **L)** Predicted percentage of cells for each T cell type; **M)** Normalized average expression for selected markers in the reference map, in WT and *Tox* KO cells.

As a second example of projection, we applied ProjecTILs to study the effect of deleting the transcription factor *Tox* in virus-specific CD8 T cells during chronic viral infection. Projection of the data from Yao *et al*.[39] revealed that *Tox*-KO cells had a dramatic alteration in state composition compared to WT controls (Figure 4K). In particular, it showed a large increase in the fraction of SLECs, at the expense of Memory Precursors and Precursor Exhausted cells (Figure 4L). Analysis of markers profiles confirmed that the *Tox*-KO cells classified as SLEC express high levels of *Klrg1* (Figure 4M). These results are consistent with previous studies that demonstrated that TOX is required for the establishment of the exhaustion CD8 T cell program[26,39,43–45].

Finally, we projected P14 cells from CD4-depleted (or isotype control) chronically infected mice from Kanev *et al*.[46]. We observed that anti-CD4 antibody-treated mice had a dramatic shift in their T cell state composition from short-lived effectors (SLEC) to Precursor Exhausted (Tpex) compared to control mice, while the proportion of Tex remained similar (Suppl. Figure S10A-B). This effect is in agreement with the conclusions of the original study, which found that in chronic infections, progenitor cells (Tpex) are unable to properly differentiate into effector cells in the absence of CD4 T cell help[46], and therefore “accumulate” in this state. We could also confirm that CD4 depletion had an effect on increasing *Pdcd1* expression, especially in Tex cells (Suppl. Figure S10C). These data were generated using the SCRB-seq protocol[47], a type of sequencing that was not included in the reference LCMV atlas. Therefore, this example also highlights the robustness of ProjecTILs to accurately project, and enable the correct interpretation of, single-cell data across multiple sequencing platforms.

## Discussion

We share the goal of many others to be able to "read" immunological states in health and disease by single-cell transcriptomics and identify therapeutic opportunities. Definition of robust, biologically relevant cellular states by scRNA-seq analysis is typically an iterative, time-consuming process that requires advanced bioinformatics and biological domain expertise. Even after successful analysis, cell clusters are not directly comparable between studies, preventing us from learning general biological rules across cohorts, conditions and models. Construction of single-cell transcriptomic atlases is a very effective approach to condense the diversity of molecular profiles within

a cell type, a tissue, or an entire organism. This is particularly powerful to characterize the heterogeneity of cell populations, to elucidate mechanisms and trajectories of differentiation as well as to aid the design of therapies targeted to specific cell types. Reference atlases can also serve as a reliable, stable baseline for the interpretation of new experiments, against which to evaluate the effect of cellular perturbations, such as changes in the balance between pre-existing states or the identification of novel states in response to immunotherapies. In this work we described a novel computational method to interpret immunological states by projecting scRNA-seq data onto a reference T cell atlas, allowing the analysis of query samples in the context of a stable, curated collection of T cell states.

To be useful, reference atlases must be detailed enough to describe all previously identified states of each cell type, and offer a level of resolution that allows discriminating biologically relevant differences between cell states. While whole-organism atlases provide an important overview on the complexity of cell populations at a global scale, the definition of high-resolution transcriptional spaces requires single-cell atlases constructed for individual cell types. Ultimately, the boundaries of a reference map should be determined by the transcriptional plasticity of individual cells in a given biological context, to avoid as much as possible distorting the reference space with "impossible" transcriptional states (for instance, a transcriptional space between T cells and fibroblasts). On this premise, useful whole-organism atlases could be described as a collection of disjoint transcriptional spaces defined independently, at high-resolution, for separate cell types (developmental studies would represent exceptional cases).

The reference atlases presented here, describe the transcriptional heterogeneity of murine T cells in cancer and viral infection, covering two of the most studied models of T cell response, in particular to identify the mechanism of action of cancer immunotherapies. While we observed that the main, known T cell subtypes can be recapitulated in these reference maps, it is likely that these atlases do not yet encompass the full diversity of transcriptional states that can be acquired by T cells, especially for CD4 TILs (which were under-represented compared to CD8 among available data) and for $\gamma\delta$ T cells (which were not represented). Only very recently, with the popularization of single-cell technologies, it has become possible to generate data of sufficient depth and quality to construct such high-resolution reference maps. We are therefore just beginning to appreciate the full potential of combining information from multiple studies and perform meta-analyses across models, tissues and cancer types. Considering the pace at which new single-cell data are generated, we anticipate that reference maps will quickly grow in size and completeness, increasingly covering the space of possible transcriptional states that can be assumed by individual cells. We expect that projection of new data into such exhaustive reference atlases will also improve as a consequence.

A fundamental level of analysis at single-cell resolution is the identification of cell subtypes or states. While it is useful, for the design of experimental validations, to summarize cell subtypes/states into simple molecular signatures composed of a few marker genes, it is also apparent that cells exist in a continuum of states, which

would be best described as probability distributions, or regions in a multi-dimensional transcriptional space. A key advantage of embedding new data into a reference atlas of cell states is that the query cells can be interpreted in a continuous space of transcriptional states, allowing the visualization of dynamic changes between *e.g.* experimental conditions. While ProjecTILs can be thought of as a classifier into pre-annotated states, it also operates in a continuous space and can, therefore, capture intermediate and transient cellular states. For instance, we observed a pattern of expression of the chemokine receptor *Cx3cr1* that straddles the Tpex and Tex subsets in the TIL atlas (Figure 1F), as well as a gradient of expression for this molecule within the Tex compartment in infection. Indeed, recent studies have identified *Cx3cr1* as a marker for a transitory, intermediate state between the Tpex and Tex populations[48–50].

Previously developed methods such as *scmap*[51], based on nearest neighbor search, allow 'matching' query cells to annotated cells or clusters that are part of a reference dataset, enabling the classification of new data into predefined cell types. Beyond the interpretation of new data in terms of known, annotated cell states, we have shown that ProjecTILs can aid the discovery of novel states that deviate from the reference as a consequence of *e.g.* a genetic alteration. For example, analysis of the Regnase-1 KO data of Wei and colleagues not only explained the increased T cell persistence and tumor control in terms of previously known T cell states (Figure 3E) but also revealed the previously unreported down-regulation of a novel inhibitory gene program in exhausted CD8 TILs. This program (ICA 25) was driven by *Klrc1*/NKG2A and *Lag3* (Figure 3G-H), and its expression was uncoupled from that of other inhibitory receptors such as PD-1, TIM-3 or CTLA4, suggesting a potential benefit of targeting these two programs simultaneously. Indeed, dual blockade of PD-1 and LAG-3 results in robust and synergistic reinvigoration of Tex cells in cancer[52] and chronic infection[53], while NKG2A blockade has been recently shown to potentiate CD8 T cell immunity induced by cancer vaccines[54].

In most experiments aimed at studying the effect of a perturbation, it is imperative to design a control group, as a baseline to compare the effect of the perturbation of interest. In single-cell experiments, when studying the effect of a genetic perturbation, or of a treatment, the control group usually consists of the same cell population as the perturbation group but under basal conditions. It is conceivable, as reference atlases become increasingly complete, that the control group is already satisfactorily included in the reference atlas – the new condition could then be directly evaluated against the atlas, bypassing the need for deeply re-sampling the transcriptional space of basal conditions, as we have illustrated with the analysis of Regnase-1 KO data in absence of a control sample (Suppl. Figure S9).

We have described the construction of reference single-cell atlases for murine T cells in pan-cancer and infection models that are strongly supported by literature. Compared to murine models, the large genetic and environmental variability, tumor heterogeneity and differences in prior antigenic impact (among other factors) between patients, present additional challenges to the analysis of human T cell phenotypes. Nonetheless, single-cell transcriptomes consisting of thousands of tumor-infiltrating T cells are also increasingly becoming available

from patients[55–61]. We envision that soon it will become feasible to construct high-quality reference maps for human T cell states, and interpret new experiments by projecting the data onto human T cell atlases, in the same way as we have shown for mouse. Moreover, it may become possible to combine murine and human maps over a layer of homologous genes, in this way leveraging information – antigenic specificity in particular – and biological insight from the murine to the human system. If T cell receptor (TCR) sequences are available for single cells in a query set, interesting questions about clonal immunodominance can be addressed by tracking individual clones over low-dimensional representations of the reference atlas. Finally, projecting whole tissue data into a collection of cell type atlases, would enable interpreting immune responses at a higher level of complexity, by the study of correlation, and putative interaction and cross-talk, between cell types at very specific cellular states.

We have implemented ProjecTILs as an R package (https://github.com/carmonalab/ProjecTILs) and we provide a Docker image ready to use. Because ProjecTILs is integrated with Seurat3, it can be easily combined with other tools for up- and down-stream analyses. We and our collaborators are extensively using ProjecTILs to reveal how experimental cancer therapies modulate TIL states, and we believe this approach will have a great impact in revealing the mechanisms of action of experimental immunotherapies and guiding novel therapeutic interventions in cancer and beyond.

## Methods

### Single-cell RNA-seq of tumor-draining lymph node T cells

Tumor draining lymph nodes from mice with established MC38 tumors (~190 $mm^3$) were excised and single-cell suspensions were stained and sorted into $CD45^+TCRb^+$ T cells. Sorted cells were then loaded onto a 10x Chromium Chip A using reagents from the Chromium Single-Cell 5' Library and Gel Bead Kit (10x Genomics) according to the manufacturer's protocol. Amplified cDNA was used for both 5' RNA-seq library generation and TCR V(D)J targeted enrichment using the Chromium Single-Cell V(D)J Enrichment Kit for Mouse T Cells (10x Genomics). 5' RNA-seq and TCR V(D)J libraries were prepared following the manufacturer's user guide (10x Genomics). The final libraries were profiled using the Bioanalyzer High Sensitivity DNA Kit (Agilent Technologies) and quantified using the Kapa Library Quantification Kit (Kapa Biosystems). Each single-cell RNA-seq library was sequenced in one lane of HiSeq4000 (Illumina) to obtain a minimum of 20,000 paired-end reads (26 x 98 bp) per cell. Single-cell TCR V (D)J libraries were multiplexed and sequenced in one lane of HiSeq2500 (Illumina) to obtain minimum of 5,000 paired-end reads (150 x 150 bp) per cell. The sequencing specifications for both single-cell RNA-seq and TCR V(D)J libraries were according to the manufacturer's specification (10x Genomics).

Single-cell RNA-seq data for each replicate were processed using cellranger *count* [CellRanger 2.2.0 (10x Genomics)] using a custom reference package based on mouse reference genome GRCm38 and GENCODE[62] gene models. Individual count tables were merged using CellRanger *aggr* to reduce batch effects.

**Public Single-cell RNA-seq data**

To construct the reference TIL atlas, we obtained single-cell gene expression matrices from the following GEO entries: GSE124691[63], GSE116390[25], GSE121478[33], GSE86028[64]; and entry E-MTAB-7919[65] from Array-Express. Data from GSE124691 contained samples from tumor and from tumor-draining lymph nodes, and were therefore treated as two separate datasets. For the TIL projection examples (OVA Tet+, miR-155 KO and Regnase-KO), we obtained the gene expression counts from entries GSE122713[16], GSE121478[33] and GSE137015[38], respectively.

Single-cell data to build the LCMV-specific CD8 T cell reference map were downloaded from GEO under the following entries: GSE131535[41], GSE134139[40] and GSE119943[39], selecting only samples in wild type conditions. Data for the Ptpn2-KO, Tox-KO and CD4-depletion projections were obtained from entries GSE134139[40], GSE119943[66], and GSE137007[46] and were not included in the construction of the reference map.

**Batch-effect correction and construction of reference atlases**

Prior to dataset integration, single-cell data from individual studies were filtered using TILPRED-1.0 (https://github.com/carmonalab/TILPRED), which removes cells not enriched in T cell markers (e.g. *Cd2*, *Cd3d*, *Cd3e*, *Cd3g*, *Cd4*, *Cd8a*, *Cd8b1*) and cells enriched in non-T cell genes (e.g. *Spi1*, *Fcer1g*, *Csf1r*, *Cd19*). Dataset integration was performed using STACAS[23] (https://github.com/carmonalab/STACAS), a batch-correction algorithm based on Seurat 3[67]. For the TIL reference map, we specified 600 variable genes per dataset, excluding cell cycling genes, mitochondrial, ribosomal and non-coding genes, as well as genes expressed in less than 0.1% or more than 90% of the cells of a given dataset. For integration, a total of 800 variable genes were derived as the intersection of the 600 variable genes of individual datasets, prioritizing genes found in multiple datasets and, in case of draws, those derived from the largest datasets. We determined pairwise dataset anchors using STACAS with default parameters, and filtered anchors using an anchor score threshold of 0.8. Integration was performed using the *IntegrateData* function in Seurat3, providing the anchor set determined by STACAS, and a custom integration tree to initiate alignment from the largest and most heterogeneous datasets. Similarly, to construct the LCMV reference map, we split the datasets into five batches that displayed strong technical differences, and applied STACAS to mitigate their confounding effects. We computed 800 variable genes per batch, excluding cell cycling genes, ribosomal and mitochondrial genes, and computed pairwise anchors using 200 integration genes, and otherwise default STACAS parameters. Anchors were filtered at the default threshold 0.8 percentile, and integration was performed with the *IntegrateData* Seurat3 function with the guide tree suggested by STACAS.

Both for the TIL and LCMV atlases, we performed unsupervised clustering of the integrated cell embeddings using the Shared Nearest Neighbor (SNN) clustering method[68] implemented in Seurat 3 with parameters {*resolution*=0.6, *reduction*="umap", *k.param*=20} for the TIL atlas and {*resolution*=0.4, *reduction*="pca", *k.param*=20} for the LCMV atlas. We then manually annotated individual clusters (merging clusters when necessary) based on several criteria: i) average expression of key marker genes in individual clusters; ii) gradients of gene expression over the UMAP representation of the reference map; iii) gene-set enrichment analysis to determine over- and under- expressed

genes per cluster using MAST[69]. In order to have access to predictive methods for UMAP, we recomputed PCA and UMAP embeddings independently of Seurat3 using respectively the *prcomp* function from basic R package "stats", and the "umap" R package (https://github.com/tkonopka/umap).

**The ProjecTILs pipeline**

The essential input to the ProjecTILs pipeline is an expression matrix, where genes are rows and cells are columns. If raw counts (e.g. UMI counts) are provided, each entry *x* in the matrix will be normalized using the formula: log (1 + 10,000 x / S), where S is the sum of all counts for that cell, and log is the natural logarithm. To ensure that only T cells are included in the query dataset, by default TILPRED-1.0 is applied to predict the composition of the query, and all cells annotated as "Non-T cells" or "unknown" are removed from the query. This filter can be optionally disabled by the user. Then, a reference atlas of annotated cells states (by default the TIL atlas) is loaded into memory, together with its cell embeddings in gene, PCA and UMAP spaces, and all associated metadata. In order to bring the query data in the same representation spaces as the reference map, batch-effect correction is applied to the normalized cell-gene counts of the query set using the anchor-finding and integration algorithms implemented in STACAS and Seurat3, where the genes for integration consist of the intersection of the variable genes of the reference map and all genes from the query. After batch-effect correction, the PCA rotation matrix pre-calculated on the reference atlas (i.e. the coefficients allowing the transformation from reference gene space into PCA space) is applied to the normalized, batch-corrected query matrix. In the same way, the *predict* function of the "umap" package allows transforming PCA embeddings into UMAP coordinates. By these means, the query data can be embedded into the original, unaltered coordinate spaces of the reference atlas, enabling joint visualization as well as classification of the query cells into T cell states.

ProjecTILs is implemented as a modular R package, with several functions that aid interpretation and analysis. The *make.projection* function is the core utility that implements the projection algorithm described above. It can be run in "direct" mode, in which case the PCA and UMAP rotations are directly applied without batch-effect correction. This may be useful for very small datasets, where alignment and integration algorithms will not be applicable. *Plot.projection* allows visualizing the query dataset as density level curves superimposed on the reference atlas. The *cellstate.predict* function implements a nearest-neighbor classifier, which predicts the state of each query cell by a majority vote of its annotated nearest neighbors (either in PCA or UMAP space) in the reference map. *Find.discriminant.dimensions* analyses PCA and ICA embeddings (described below) to determine dimensions where the query deviates significantly from the reference map. Several additional functions allow visualizing multiple aspects of the reference and projected dataset and aid the biological interpretation of the results. The code and description of the package can be found at: https://github.com/carmonalab/ProjecTILs

**ICA and discriminant dimensions**

Independent component analysis (ICA) is a computational technique aimed at deconvoluting a multivariate signal (such as simultaneous expression of many genes) into additive, independent sources (in this case different genetic programs). We applied the fastICA implementation[70] to determine 50 independent components on the integrated expression matrix of the TIL reference atlas. To suggest a biological interpretation of the ICA components, we downloaded hallmark gene sets (H) and canonical pathway gene sets (CP) from the Molecular Signatures Database[37] (mSigDB), as well as selected immunological signatures from previous studies. We scored each ICA against these signatures, retaining the top three-scoring signatures for each ICA, and clustered ICA components based on the union of all retained signatures (see Suppl. Figure S6).

After projection, query datasets are also subject to transformation in ICA space through the ICA rotation matrix. The *find.discriminant.dimensions* function implemented in ProjecTILs evaluates the distribution of the cells in the query for each ICA dimension, and compares it to the distribution of cells in the reference (or to a control query dataset, if provided) in the same ICA dimension. For each ICA dimension, a statistical test can then be applied to confirm or reject the null hypothesis that, in this dimension, the cells in reference and query are drawn from the same distribution. ProjecTILs implements a Kolmogorov-Smirnov (KS) test or a t-test to the null hypothesis, multiplying p-values by the number of tests (i.e. 50) to correct for multiple testing (*i.e.* Bonferroni correction). ICA dimensions where the query deviates significantly from the reference (or the control, if provided) are ranked by their test statistic to determine the top discriminant dimensions. ICA embeddings can be visualized as an additional dimension on the z-axis of the reference UMAP space with the function *plot.discriminant.3d*.

**Cross-validated projection benchmark**

To estimate the accuracy of the projection algorithm, we devised a cross-validation algorithm where we removed part of the data from the reference map before projecting it back into the map. Each of the seven datasets included in the reference TIL map, except one, was composed of at least two samples; we constructed a cross-validation experiment by removing, at each cross-validation step, half of the samples of a given dataset, and then projected these samples into a reduced version of the map that does not contain the data points from these samples. The dataset by Singer *et al.*[64] consists of a single sample, and therefore we removed all of its cells before projecting it back into its reduced map. After systematically projecting all cells in cross-validation, we compared their projected coordinates (either in UMAP or PCA space) with their original coordinates in the reference map. By the same token, we evaluated the performance of the projection algorithm by comparing the composition of any given dataset in terms of predicted cell states compared to the annotated cell states in the reference map (confusion matrix, Figure 2G). We performed cross-validation for the complete projecTILs algorithm using default parameters, as well as to the "direct" mode of projection, which bypasses batch-effect correction and applies directly the PCA and UMAP transformations on unaligned gene-expression data. We found that using ProjecTILs with batch-correction, 90.0% of the projected cells are found within a radius of one unit (their neighborhood) in

UMAP space (Figure 2H, Suppl. Figure S1) from their original coordinate in the reference map, and 93.4% within a radius of 5 units in PCA space. In contrast, direct projection (i.e. without alignment) places 59.7% of cells within a radius of one unit in UMAP space, and 62.3% within a radius of 5 units in PCA space (Suppl. Figure S5). While batch-correction has a clear positive impact also in terms of predicted T-cell state composition (Suppl. Figure S2, S4), direct projection is a viable alternative with reasonable performance, especially in the case of very small query datasets, where batch correction algorithms may not be applicable.

## Accession Codes

Generated scRNA-seq data of MC38 tumor-draining lymph node T cells (excluding antigen-inexperienced CD62L+CD44- cells) were deposited in the ArrayExpress database with accession ID E-MTAB-9274.
Published scRNA-seq data of MC38 TILs: E-MTAB-7919, GSE124691
Published scRNA-seq data of B16 TILs GSE116390, GSE121478, GSE86028, GSE122713, GSE137015
Published scRNA-seq data of LCMV infection P14 cells: GSE131535, GSE134139, GSE119943, GSE119943, GSE137007

## Conflicts of interest

MA, JCO, SJC declare no conflicts of interest.
GC has received grants, research support or is coinvestigator in clinical trials by BMS, Celgene, Boehringer Ingelheim, Roche, Iovance and Kite; has received honoraria for consultations or presentations by Roche, Genentech, BMS, AstraZeneca, Sanofi-Aventis, Nextcure and GeneosTx; he has patents in the domain of antibodies and vaccines targeting the tumor vasculature as well as technologies related to T-cell expansion and engineering for T-cell therapy; and he receives royalties from the University of Pennsylvania for CAR-T technologies licensed to Novartis.
SM and RC are employees of Genentech, Inc, a member of the Roche family and receive salary and stock from Roche.

## Acknowledgements

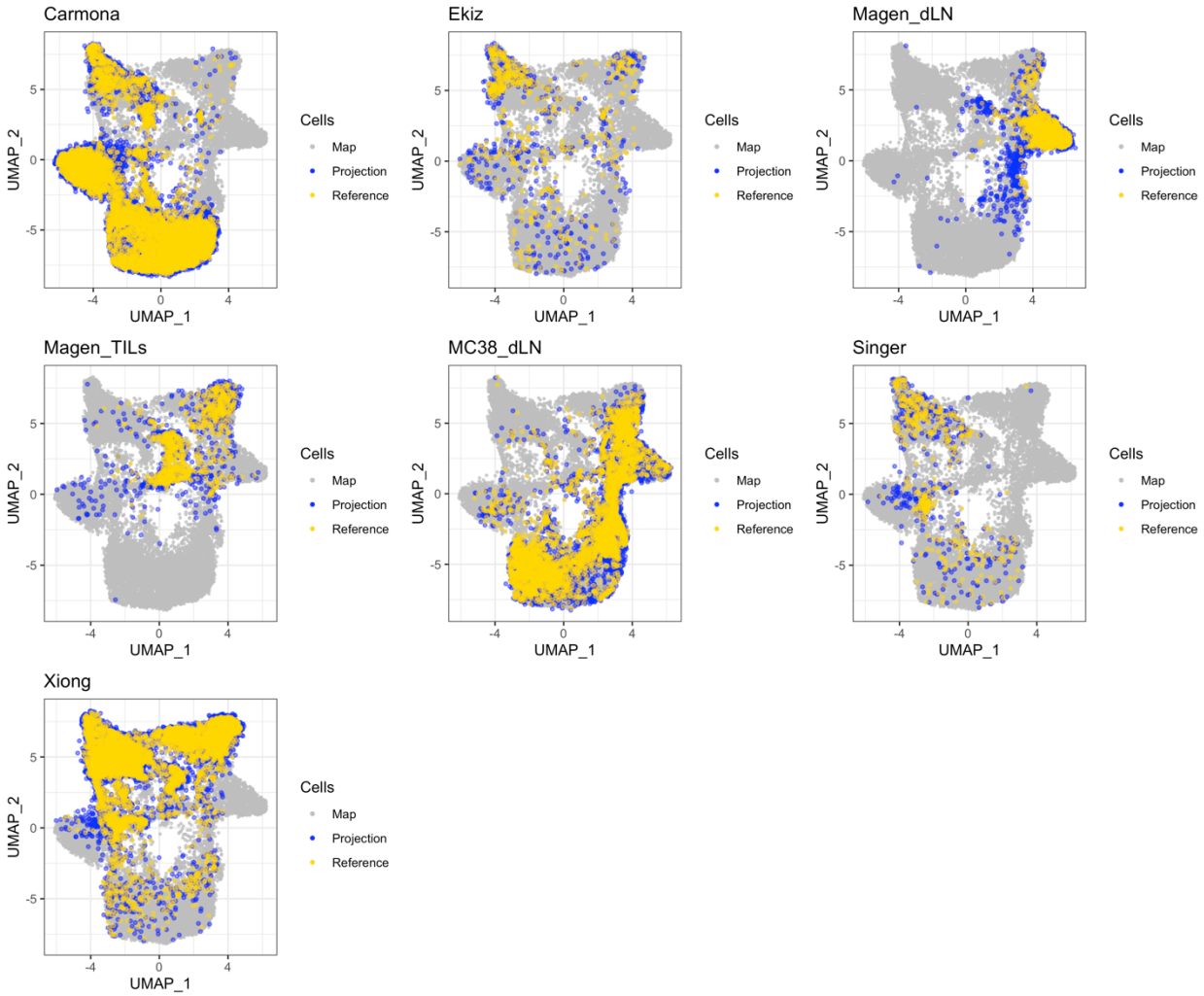
We are grateful to Dr. Amaia Martinez-Usatorre and Dr. Gonzalo Parra for critical reading of the manuscript. This research was possible thanks to the support of the Swiss National Science Foundation (SNF) Ambizione grant 180010 to SJC. SJC would like to thank the SNF for supporting researchers developing their early independent academic careers in Switzerland, regardless of their nationalities.

## References

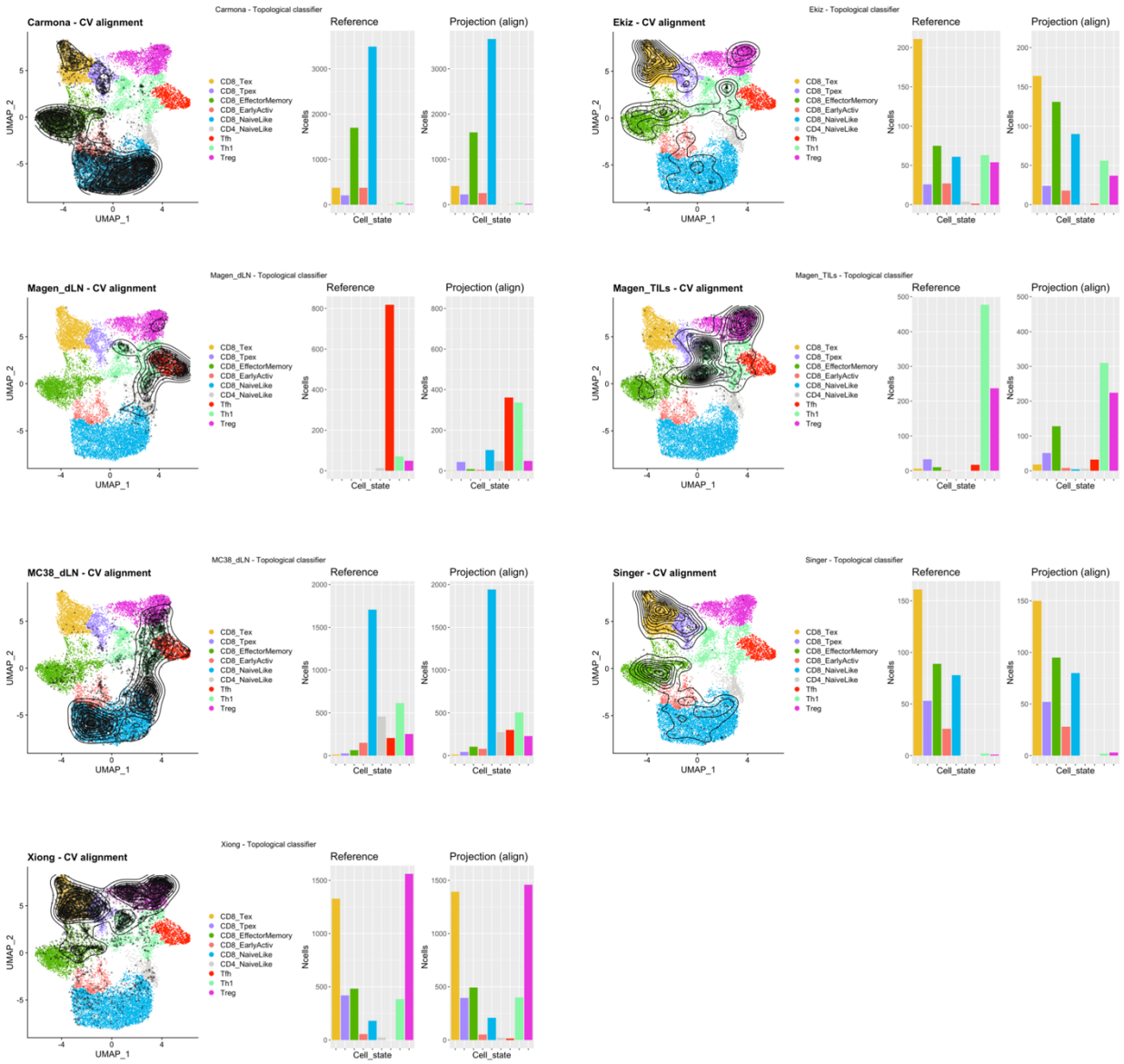

1. Efremova, M., Vento-Tormo, R., Park, J.-E., Teichmann, S. A. & James, K. R. Immunology in the Era of Single-Cell Technologies. *Annu. Rev. Immunol.* **38**, 727–757 (2020).
2. Yofe, I., Dahan, R. & Amit, I. Single-cell genomic approaches for developing the next generation of immunotherapies. *Nat. Med.* **26**, 171–177 (2020).
3. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, (2017).
4. van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+ T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
5. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12**, 298–306 (2012).
6. Tumeh, P. C. *et al.* PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* (2014) doi:10.1038/nature13954.
7. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293–1308.e36 (2018).
8. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998–1013.e20 (2018).
9. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775–789.e18 (2019).
10. Blank, C. U. *et al.* Defining 'T cell exhaustion'. *Nat. Rev. Immunol.* 1–10 (2019) doi:10.1038/s41577-019-0221-9.
11. Paley, M. A. *et al.* Progenitor and terminal subsets of CD8+ T cells cooperate to contain chronic viral infection. *Science* **338**, 1220–1225 (2012).
12. Utzschneider *et al.* T Cell Factor 1-Expressing Memory-like CD8+ T Cells Sustain the Immune Response to Chronic Viral Infections. *Immunity* **45**, 415–427 (2016).
13. Wu, T. *et al.* The TCF1-Bcl6 axis counteracts type I interferon to repress exhaustion and maintain T cell stemness. *Sci. Immunol.* **1**, (2016).
14. Im, S. J. *et al.* Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
15. 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.e10 (2019).
16. Miller, B. C. *et al.* Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade. *Nat. Immunol.* **20**, 326–336 (2019).
17. Brummelman, J. *et al.* High-dimensional single cell analysis identifies stem-like cytotoxic CD8+ T cells infiltrating human tumors. *J. Exp. Med.* **215**, 2520–2535 (2018).
18. Kim, H.-J. & Cantor, H. CD4 T-cell subsets and tumor immunity: the helpful and the not-so-helpful. *Cancer Immunol. Res.* **2**, 91–98 (2014).
19. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
20. Johansen, N. & Quon, G. scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data. *Genome Biol.* **20**, 166 (2019).
21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
22. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
23. Andreatta, M. & Carmona, S. J. STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data. *bioRxiv* 2020.06.15.152306 (2020) doi:10.1101/2020.06.15.152306.
24. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
25. Carmona, S. J., Siddiqui, I., Bilous, M., Held, W. & Gfeller, D. Deciphering the transcriptomic landscape of tumor-infiltrating CD8 lymphocytes in B16 melanoma tumors with single-cell RNA-Seq. *OncoImmunology* **9**, (2020).
26. Scott, A. C. *et al.* TOX is a critical regulator of tumour-specific T cell differentiation. *Nature* **571**, 270–274 (2019).
27. Kallies, A., Zehn, D. & Utzschneider, D. T. Precursor exhausted T cells: key to successful immunotherapy? *Nat. Rev. Immunol.* 1–9 (2019) doi:10.1038/s41577-019-0223-7.
28. Held, W., Siddiqui, I., Schaeuble, K. & Speiser, D. E. Intratumoral CD8+ T cells with stem cell-like properties: Implications for cancer immunotherapy. *Sci. Transl. Med.* **11**, (2019).
29. Magen, A. *et al.* Single-Cell Profiling Defines Transcriptomic Signatures Specific to Tumor-Reactive versus Virus-Responsive CD4+ T Cells. *Cell Rep.* **29**, 3019-3032.e6 (2019).
30. Xu, W. *et al.* The Transcription Factor Tox2 Drives T Follicular Helper Cell Development via Regulating Chromatin Accessibility. *Immunity* **53**, 826–839 (2019).
31. Guillaume, P., Baumgaertner, P., Angelov, G. S., Speiser, D. & Luescher, I. F. Fluorescence-activated cell sorting and cloning of bona fide CD8+ CTL with reversible MHC-peptide and antibody Fab' conjugates. *J. Immunol. Baltim. Md 1950* **177**, 3903–3912 (2006).
32. Cebecauer, M. *et al.* Soluble MHC-peptide complexes induce rapid death of CD8+ CTL. *J. Immunol. Baltim. Md 1950* **174**, 6809–6819 (2005).
33. Ekiz, H. A. *et al.* MicroRNA-155 coordinates the immunological landscape within murine melanoma and correlates with immunity in human cancers. *JCI Insight* **4**, (2019).
34. Dudda, J. C. *et al.* MicroRNA-155 is required for effector CD8+ T cell responses to virus infection and cancer. *Immunity* **38**, 742–753 (2013).
35. Gracias, D. T. *et al.* The microRNA miR-155 controls CD8(+) T cell responses by regulating interferon signaling. *Nat. Immunol.* **14**, 593–602 (2013).
36. Engreitz, J. M., Daigle, B. J., Marshall, J. J. & Altman, R. B. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* **43**, 932–944 (2010).

37. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
38. Wei, J. *et al.* Targeting REGNASE-1 programs long-lived effector T cells for cancer therapy. *Nature* **576**, 471–476 (2019).
39. Yao, C. *et al.* Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. *Nat. Immunol.* **20**, 890–901 (2019).
40. LaFleur, M. W. *et al.* PTPN2 regulates the generation of exhausted CD8+ T cell subpopulations and restrains tumor immunity. *Nat. Immunol.* **20**, 1335–1347 (2019).
41. Chen, Z. *et al.* TCF-1-Centered Transcriptional Network Drives an Effector versus Exhausted CD8 T Cell-Fate Decision. *Immunity* **51**, 840-855.e5 (2019).
42. Manguso, R. T. *et al.* In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* **547**, 413–418 (2017).
43. Alfei, F. *et al.* TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection. *Nature* **571**, 265–269 (2019).
44. Seo, H. *et al.* TOX and TOX2 transcription factors cooperate with NR4A transcription factors to impose CD8+ T cell exhaustion. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12410–12415 (2019).
45. Khan, O. *et al.* TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion. *Nature* **571**, 211–218 (2019).
46. Kanev, K. *et al.* Proliferation-competent Tcf1+ CD8 T cells in dysfunctional populations are CD4 T cell help independent. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20070–20076 (2019).
47. Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. & Mikkelsen, T. S. *Characterization of directed differentiation by high-throughput single-cell RNA-Seq*. http://biorxiv.org/lookup/doi/10.1101/003236 (2014) doi:10.1101/003236.
48. Hudson, W. H. *et al.* Proliferating Transitory T Cells with an Effector-like Transcriptional Signature Emerge from PD-1+ Stem-like CD8+ T Cells during Chronic Infection. *Immunity* **51**, 1043-1058.e4 (2019).
49. Beltra, J.-C. *et al.* Developmental Relationships of Four Exhausted CD8+ T Cell Subsets Reveals Underlying Transcriptional and Epigenetic Landscape Control Mechanisms. *Immunity* **52**, 825-841.e8 (2020).
50. Zander, R. *et al.* CD4+ T Cell Help Is Required for the Formation of a Cytolytic CD8+ T Cell Subset that Protects against Chronic Infection and Cancer. *Immunity* **51**, 1028-1042.e4 (2019).
51. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
52. Matsuzaki, J. *et al.* Tumor-infiltrating NY-ESO-1–specific CD8+ T cells are negatively regulated by LAG-3 and PD-1 in human ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7875–7880 (2010).
53. Blackburn, S. D. *et al.* Coregulation of CD8+ T cell exhaustion by multiple inhibitory receptors during chronic viral infection. *Nat. Immunol.* **10**, 29–37 (2009).
54. van Montfoort, N. *et al.* NKG2A Blockade Potentiates CD8 T Cell Immunity Induced by Cancer Vaccines. *Cell* **175**, 1744-1755.e15 (2018).
55. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).
56. Li, H. *et al.* Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell* **176**, 775–789.e18 (2019).
57. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* **175**, 984-997.e24 (2018).
58. Zhang, L. *et al.* Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature* **564**, 268–272 (2018).
59. Guo, X. *et al.* Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
60. Thommen, D. S. *et al.* A transcriptionally and functionally distinct PD-1+ CD8+ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat. Med.* **24**, 994–1004 (2018).
61. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).
62. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
63. Magen, A. *et al.* Single-Cell Profiling Defines Transcriptomic Signatures Specific to Tumor-Reactive versus Virus-Responsive CD4+ T Cells. *Cell Rep.* (2019) doi:10.1016/j.celrep.2019.10.131.
64. Singer, M. *et al.* A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. *Cell* **166**, 1500–1511.e9 (2016).
65. Xiong, H. *et al.* Coexpression of inhibitory receptors enriches for activated and functional CD8+ T cells in murine syngeneic tumor models. *Cancer Immunol. Res.* **7**, 963–976 (2019).
66. Yao, C. *et al.* Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection. *Nat. Immunol.* **20**, 890–901 (2019).
67. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
68. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 471 (2013).
69. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* (2015) doi:10.1186/s13059-015-0844-5.
70. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Netw. Off. J. Int. Neural Netw. Soc.* **13**, 411–430 (2000).

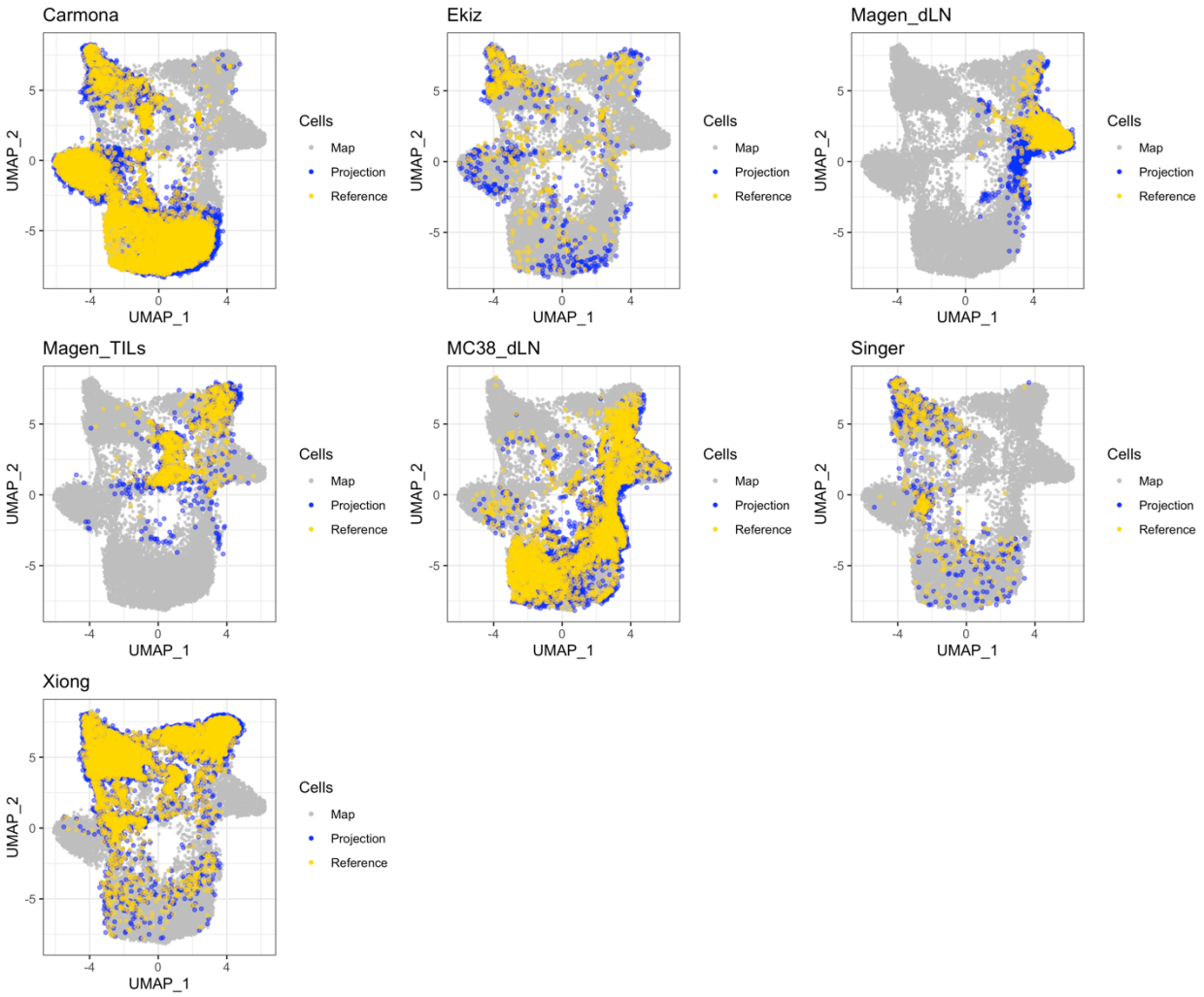
## Supplementary Material



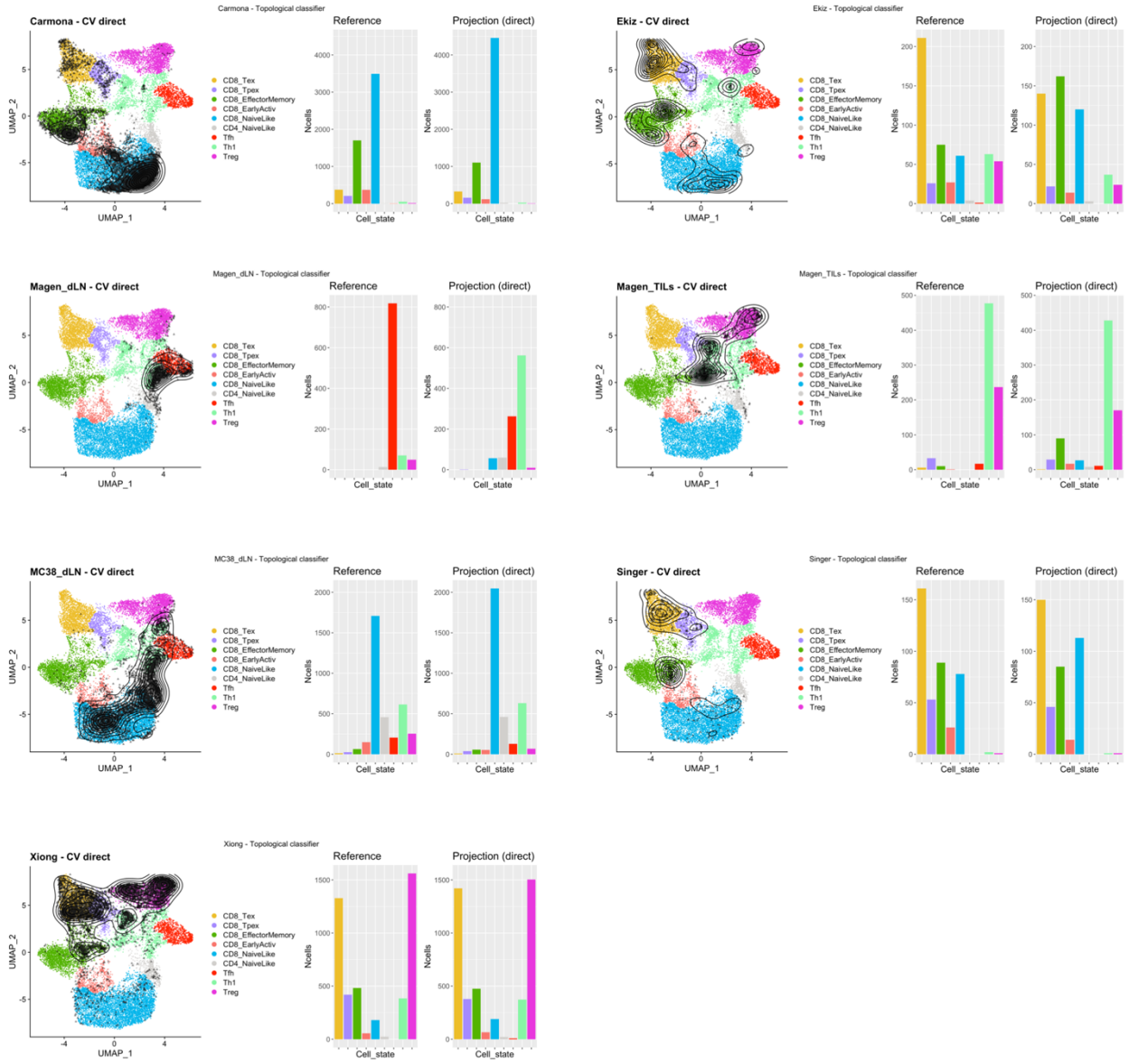

**Supplementary Figure S1: Cross-validated projection of dataset in the reference map using batch correction.** For each of the seven datasets included in the reference TIL map, we constructed a cross-validation experiment by removing, at each cross-validation step, half of the samples of a given dataset, and then projected these samples into a reduced version of the map that does not contain the data points from these samples. The projected coordinates (blue) can be compared against the original coordinates in the map (yellow); the complete map layout is shown in gray. Note that the distribution pattern of each dataset is associated to the specific T cell subsets it contains, e.g. datasets Carmona and Singer contain CD8 TILs, Magen_TILs contain CD4 TILs, Ekiz and Xiong contain both CD4 and CD8 TILs, Magen_dLN and MC38_dLN contain CD4 T cells, and both CD4 and CD8 T cells, respectively, from tumor-draining lymph nodes.



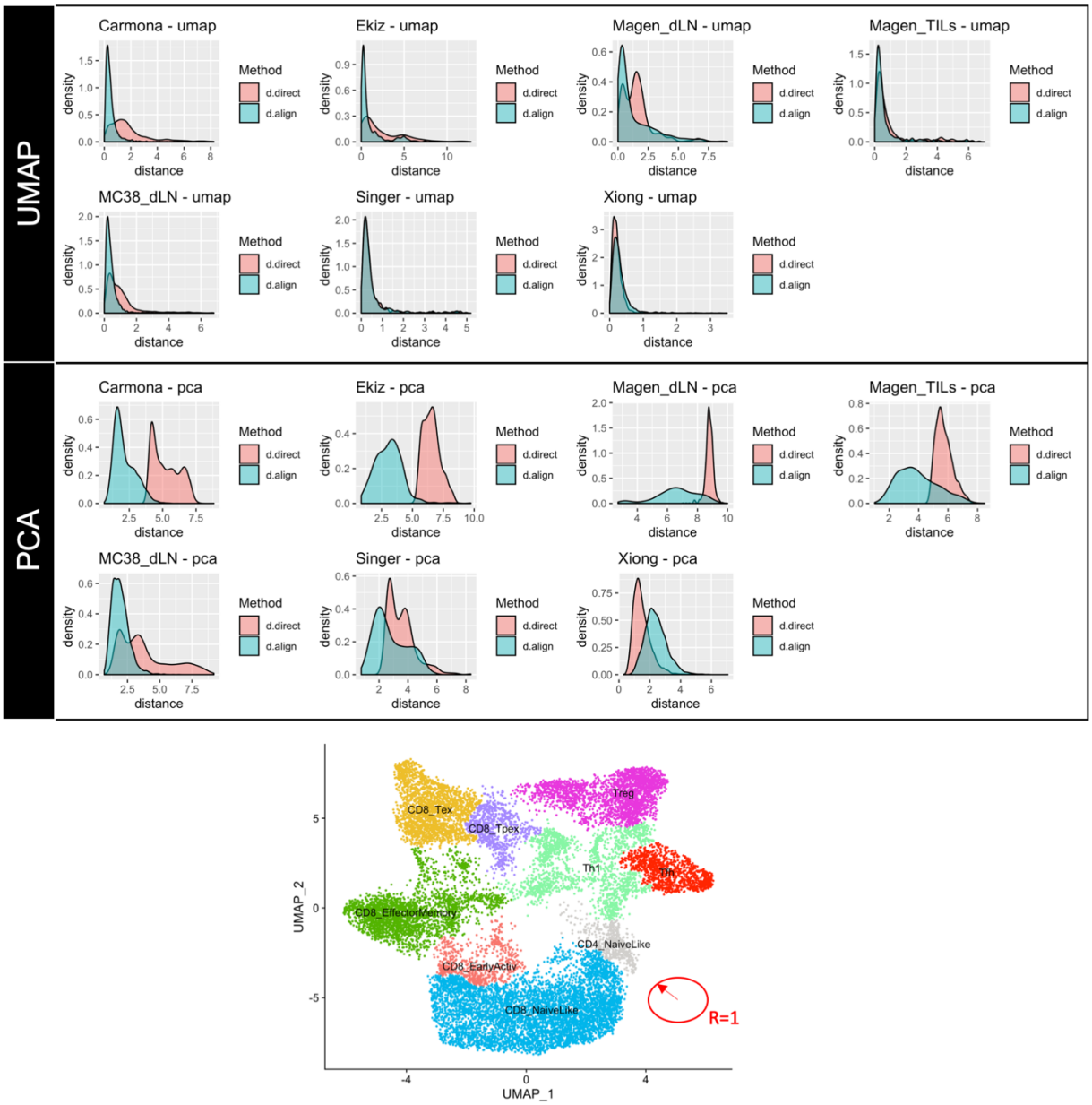

**Supplementary Figure S2: Cell state composition for reference and projection in cross-validation benchmark, using batch correction.** For each dataset, the projected coordinates in cross-validation are shown as level curves on the UMAP representation of the map (left). Based on the projected coordinates, a nearest-neighbor algorithm is applied to predict the cell state of the query cells; the predicted composition of the dataset can be then compared to the "true" composition of the reference map for the same dataset (right).



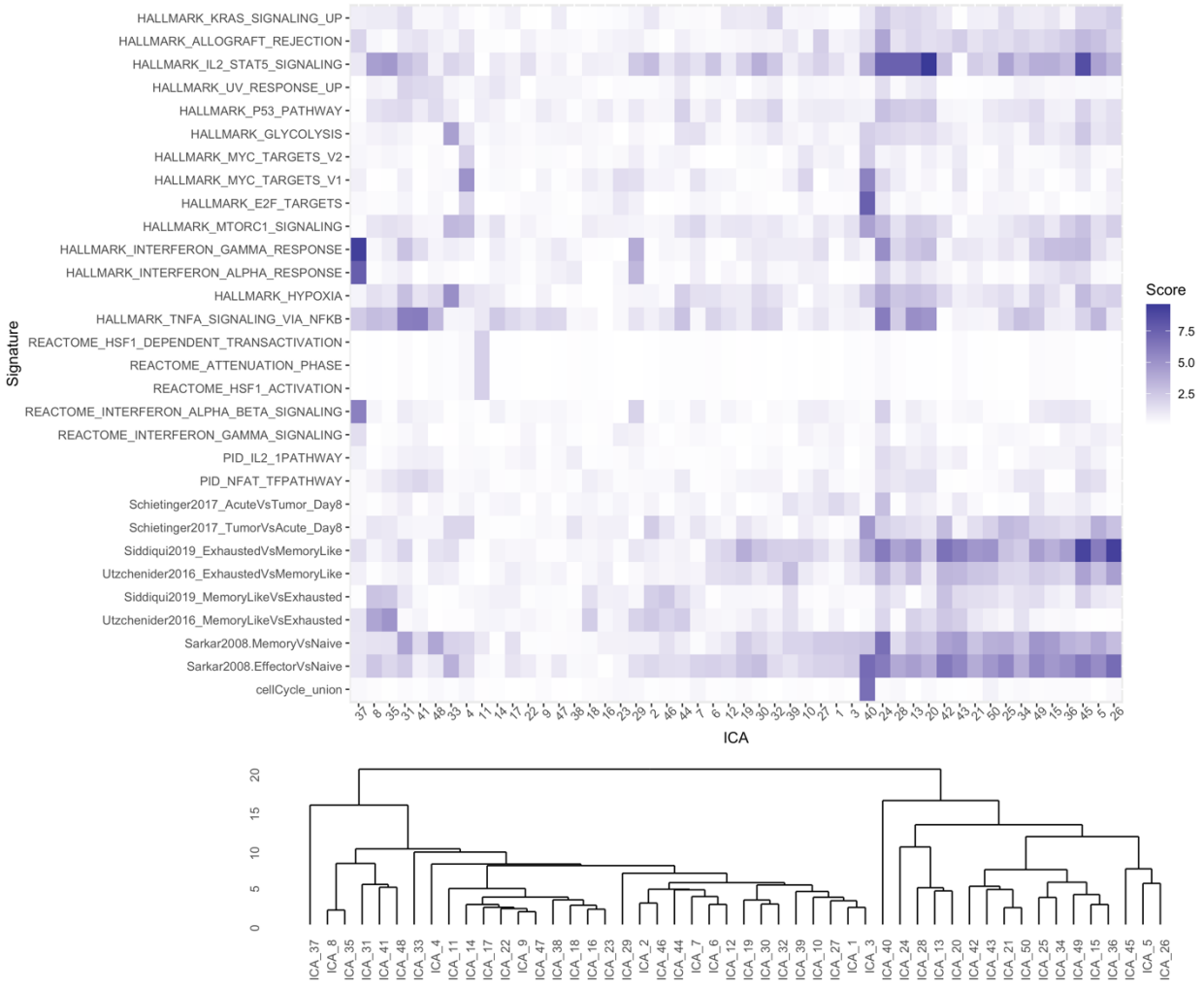

**Supplementary Figure S3: Cross-validated projection of dataset in the reference map without batch correction.** For each of the seven datasets included in the reference TIL map, we constructed a cross-validation experiment by removing, at each cross-validation step, half of the samples of a given dataset, and then directly projected these samples (excluding batch correction) into a reduced version of the map that does not contain the data points from these samples. The projected coordinates (blue) can be compared against the original coordinates in the map (yellow); the complete map layout is shown in gray.



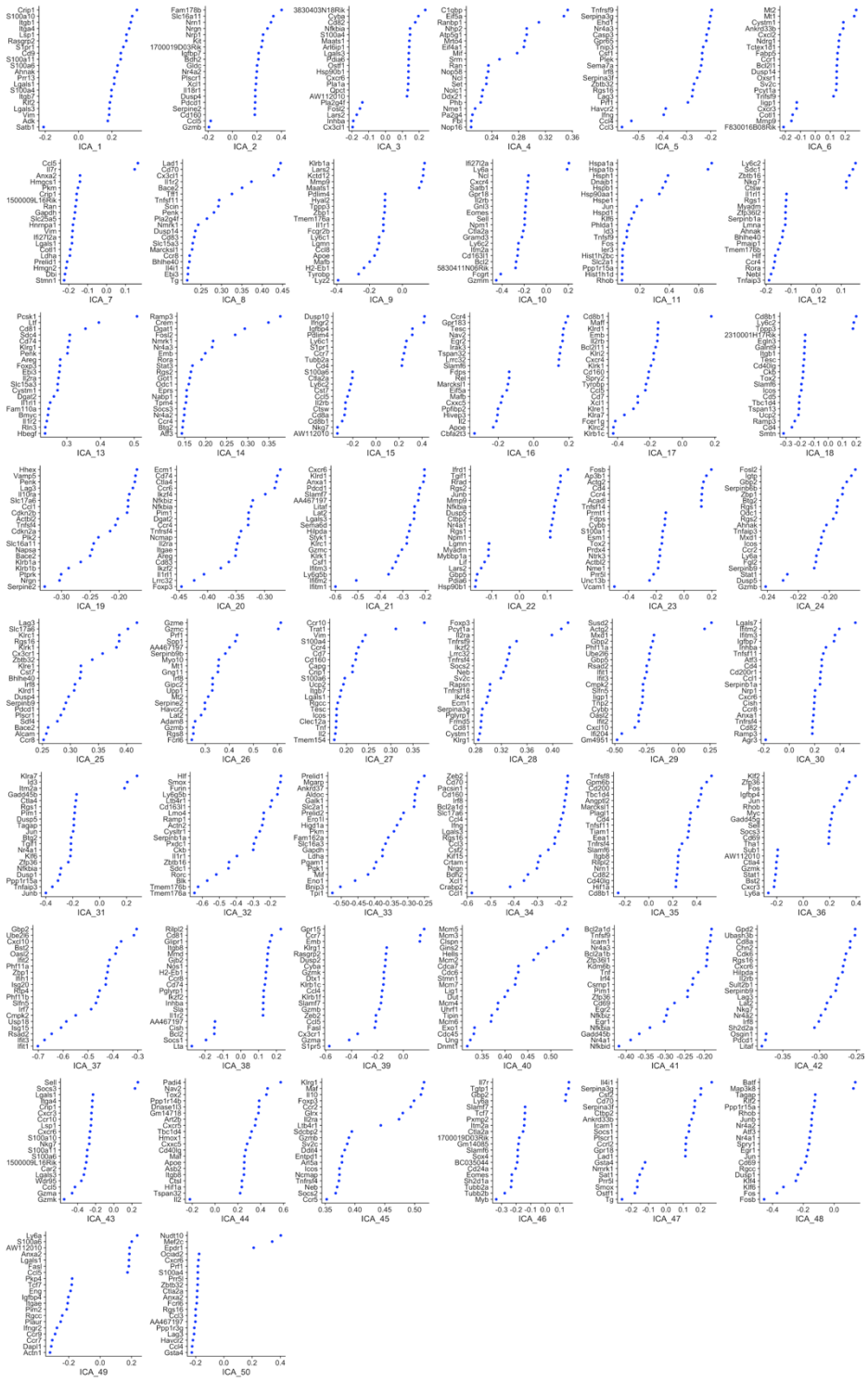

**Supplementary Figure S4: Cell state composition for reference and projection in cross-validation benchmark, without batch correction.** For each dataset, the projected coordinates in cross-validation are shown as level curves on the UMAP representation of the map (left). Based on the projected coordinates, a nearest-neighbor algorithm is applied to predict the cell state of the query cells; the predicted composition of the dataset can be then compared to the "true" composition of the reference map for the same dataset (right).



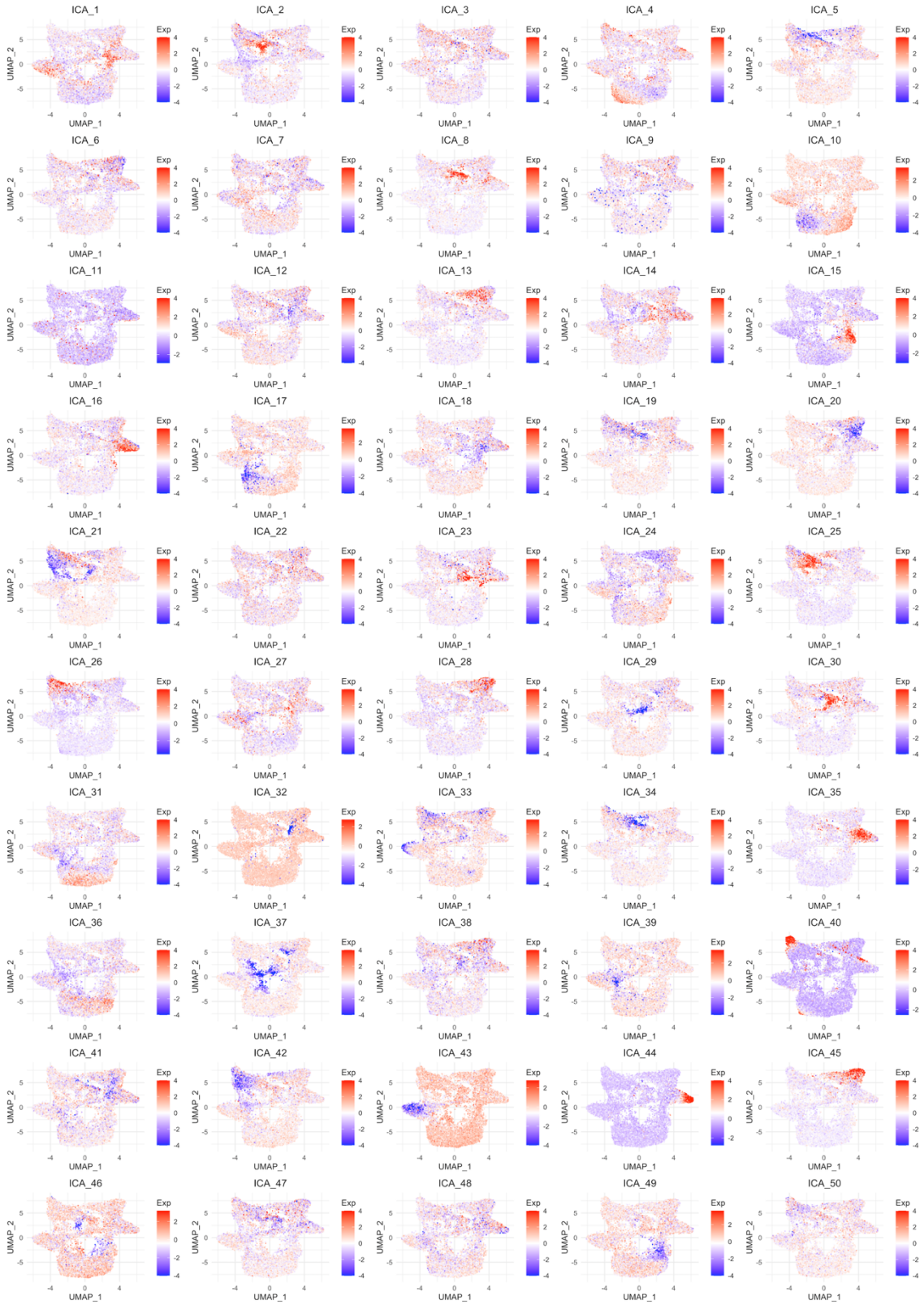

**Supplementary Figure S5: Distribution of distances between reference and predicted coordinates in cross-validated benchmark.** The reference and predicted coordinates for each cell in the map can be compared by calculating their pairwise distance in UMAP or PCA space. Density of distances for direct projection (d.direct) and batch-corrected projection (d.align) are shown for UMAP embeddings (top) and PCA embeddings (bottom). The reference map at the bottom is shown to indicate a representative radius of 1 unit in UMAP space (see main text).



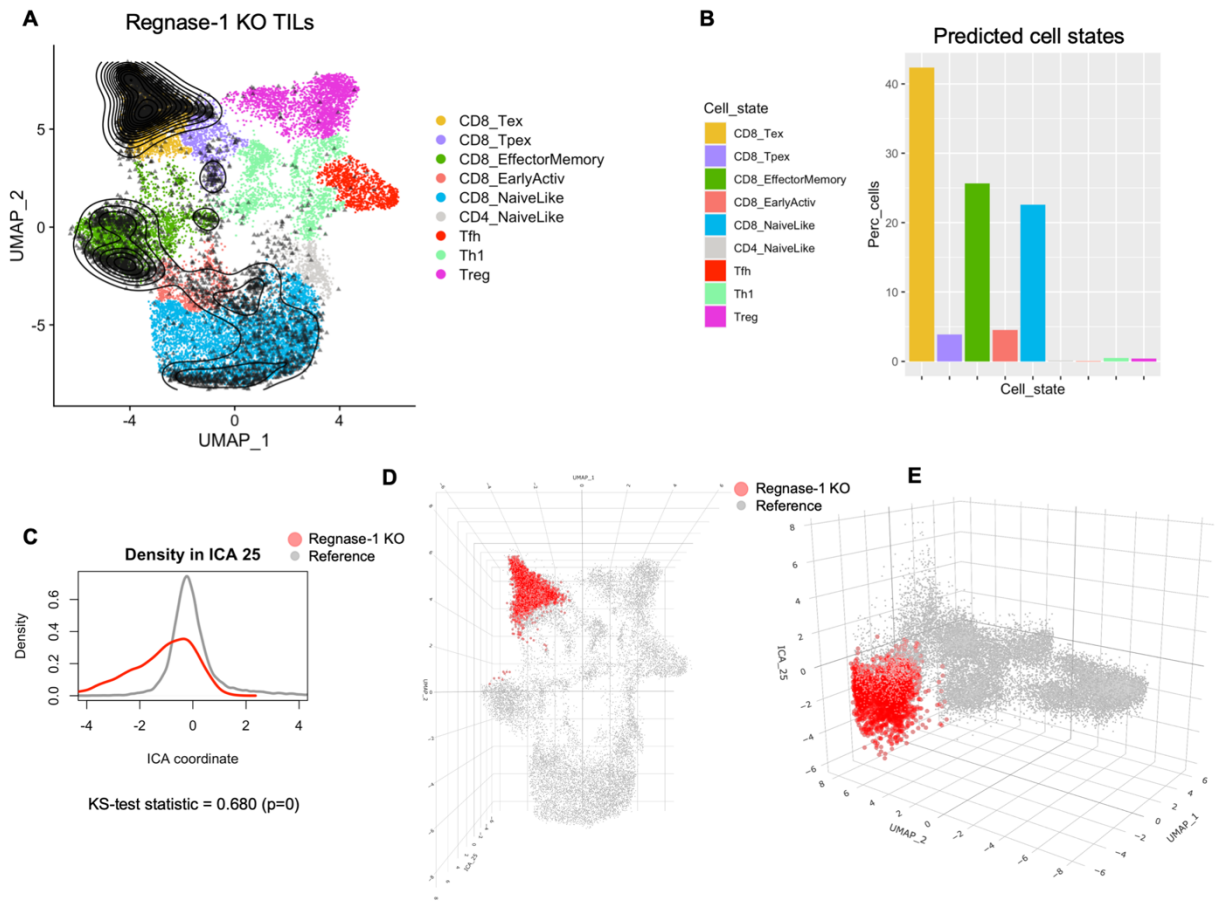

**Supplementary Figure S6: Correlation between ICA genes and annotated signatures from mSigDB.** Score for each ICA component against a selection of gene signatures from the Molecular Signature Database (top). Dendrogram of ICA components clustered by signature similarity (bottom).



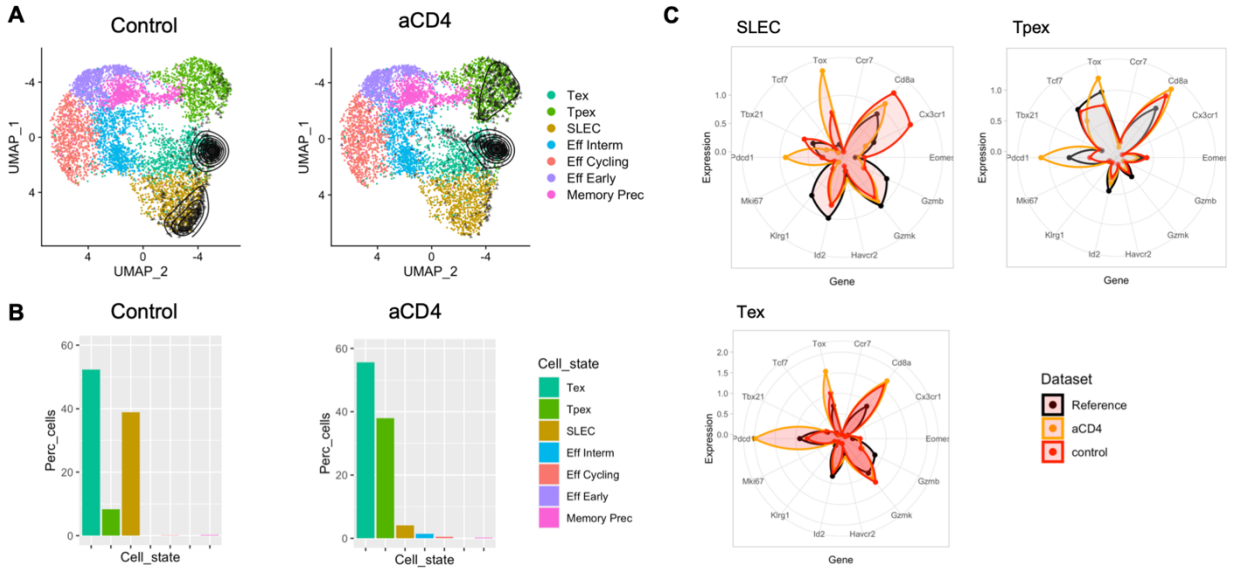
**Supplementary Figure S7:** Top 20 genes associated with each of the 50 ICA dimensions of the reference TIL atlas. The scale on the x-axis represents the ICA loading of each gene, and gives an indication of the contribution of each gene to the cell coordinate in the given ICA component.




**Supplementary Figure S8:** Reference cell embeddings in the 50 ICA dimensions of the reference TIL atlas. The Exp scale represents the coordinate of each cell for the given ICA component.




**Supplementary Figure S9: A)** Projection of Regnase-1 KO single-cell data onto the TIL reference atlas, without control sample. **B)** Predicted cell state composition in terms of nearest neighbors on the reference map. **C)** ICA component 25 was ranked as the second most significant by KS-test (statistic=0.680, p=0) for the null hypothesis that ICA 25 embeddings for CD8 exhausted Regnase-1 KO cells and CD8 exhausted reference cells come from the same distribution. **D-E)** Top and side view of (predicted) CD8 exhausted cells from Regnase-1 KO in UMAP space, with ICA 25 embeddings on the z-axis.




**Supplementary Figure S10: A)** Projection of Control and CD4-depleted (aCD4) samples from Kanev *et al.*[46] onto the LCMV reference atlas. **B)** Predicted cell state composition in terms of percentage of cells (Perc_cells) for Control and aCD4 treatment, showing a shift from short-lived effectors (SLEC) to Precursor Exhausted (Tpex) phenotype upon CD4 depletion. **C)** Normalized gene expression profile for selected marker genes for the SLEC, Tpex and Tex cell types.