Unsupervised machine learning reveals key immune cell subsets in COVID-19, rhinovirus infection, and cancer therapy

Sierra M. Barone[1,2,*], Alberta G.A. Paul[3,*], Lyndsey M. Muehling[3,4], Joanne A. Lannigan[4] William W. Kwok[5], Ronald B. Turner[6], Judith A. Woodfolk[3,4], and Jonathan M. Irish[1,2,7 ¥]

Affiliations

[1] Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA.

[2] Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA.

[3] Allergy Division, Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA, USA.

[4] Department of Microbiology, Immunology, and Cancer Biology, University of Virginia School of Medicine, Charlottesville, VA, USA.

[5] Benaroya Research Institute at Virginia Mason, Seattle, WA, USA.

[6] Department of Pediatrics, University of Virginia School of Medicine, Charlottesville, VA, USA.

[7] Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA.

[*] Denotes equal contribution.

[¥] Corresponding author and lead contact, e-mail: jonathan.irish@vanderbilt.edu (J.M.I.)

Abstract

For an emerging disease like COVID-19, systems immunology tools may quickly identify and quantitatively characterize cells associated with disease progression or clinical response. With repeated sampling, immune monitoring creates a real-time portrait of the cells reacting to a novel virus before disease specific knowledge and tools are established. However, single cell analysis tools can struggle to reveal rare cells that are under 0.1% of the population. Here, the machine learning workflow Tracking Responders Expanding (T-REX) was created to identify changes in both very rare and common cells in diverse human immune monitoring settings. T-REX identified cells that were highly similar in phenotype and localized to hotspots of significant change during rhinovirus and SARS-CoV-2 infections. MHC tetramers were not used during unsupervised analysis and instead 'left out' to serve as a test of whether T-REX identifies biologically significant cells. In the rhinovirus challenge study, T-REX identified virus-specific $CD4^+$ T cells based on these cells being a distinct phenotype that expanded by ≥95% following infection. T-REX successfully identified hotspots with virus-specific T cells using pairs of samples comparing Day 7 of infection to samples taken either after clearing the infection (Day 28) or samples taken prior to infection (Day 0). Mapping pairwise comparisons in samples according to both the direction and degree of change provided a framework to compare systems level immune changes during infectious disease or therapy response. This revealed that the magnitude and direction of systemic immune change in some COVID-19 patients was comparable to that of blast crisis acute myeloid leukemia patients undergoing induction chemotherapy and characterized the identity of the immune cells that changed the most. Other COVID-19 patients instead matched an immune trajectory like that of individuals with rhinovirus infection or melanoma patients receiving checkpoint inhibitor therapy. T-REX analysis of paired blood samples provides an approach to rapidly identify and characterize mechanistically significant cells and to place emerging diseases into a systems immunology context.

Keywords

Machine learning, rhinovirus, COVID-19, cancer immunology, systems immunology, human immune monitoring, single cell, spectral flow cytometry, mass cytometry

Introduction

Systems immunology offers a new way to compare how an individual patient's cells responds to treatment or changes during infection (Davis et al., 2017; Greenplate et al., 2016a). However, systems immunology and computational analysis tools were primarily designed to track major cell populations representing >1% of the sample. Viral immune and cancer immunotherapy responses can include mechanistically important and extremely rare T cells that proliferate rapidly over the course of days but as an aggregate exist as <0.1% of blood $CD3^+$ T cells at their peak. These cells can be tracked genetically through clonal expansion, but may be lost in computational analyses focused on describing the global landscape of phenotypes. The specific expansion or contraction of phenotypically distinct cells may be a hallmark feature of key immune effectors and could reveal these cells without the need for prior knowledge of their identity or specialized tracking reagents like MHC tetramers.

The datasets tested here were all suspension flow cytometry, a data type where it is typical to have multiple snapshot samples of cells over time, and an ongoing challenge in the field is to match or register cells to their phenotypic cognates between samples (Irish, 2014; Pyne et al., 2014; Weber and Robinson, 2016). Analysis algorithms typically rely on aggregate statistics for groups of cells, but the process of grouping the cells works best with larger, established populations (Diggins et al., 2015; Irish et al., 2006; Saeys et al., 2016) or may include pre-filtering of cells by human experts (Greenplate et al., 2016a; Greenplate et al., 2019). Cytometry tools like SPADE (Bendall et al., 2011; Qiu et al., 2011), FlowSOM (Van Gassen et al., 2015), Phenograph (Levine et al., 2015), Citrus (Bruggner et al., 2014), and RAPID (Leelatian et al., 2020) generally work best to characterize cell subsets representing >1% of the sample and are less capable of capturing extremely rare cells or subsets distinguished by only a fraction of measured features. Tools like t-SNE (Amir el et al., 2013; Krijthe et al., 2015), opt-SNE

(Belkina et al., 2019), and UMAP (Becht et al., 2018; McInnes et al., 2018) embed cells or learn a manifold and represent these transformations as algorithmically-generated axes. In addition to assisting with data visualization, these tools frequently reveal unexpected cells and facilitate their identification through manual or automated clustering (Amir el et al., 2013; Becher et al., 2014; Diggins et al., 2015; Diggins et al., 2017; Gandelman et al., 2019; Leelatian et al., 2020). Sconify (Burns et al., 2018) is one such tool that applies *k*-nearest neighbors (KNN) to calculate aggregate statistics for the immediate phenotypic neighborhood around a given cell on a t-SNE plot that combines data from multiple cytometry samples. This approach to creating a population around every cell was a key inspiration for the Tracking Responders Expanding (T-REX) tool presented here, which applies this idea to pinpoint rare cells in phenotypic regions of significant change.

Data types used to challenge the T-REX algorithm here included a new spectral flow cytometry study (Dataset 1) and three existing mass cytometry datasets (Dataset 2, Dataset 3, and Dataset 4). Mass cytometry is an established technique for human immune monitoring where commercial reagents presently allow 44 antibodies to be measured simultaneously per cell (Greenplate et al., 2016a; Mistry et al., 2019; Spitzer and Nolan, 2016). Spectral flow cytometry is gaining attention in human immune monitoring as it generates data that compares well to mass cytometry (Ferrer-Font et al., 2020; Mistry et al., 2019). Spectral flow cytometers collect cells at around 10-fold the number of cells per second as mass cytometers. While the availability of spectrally distinct antibody-fluorochrome conjugates imposes a practical limit on spectral flow cytometry at present, established panels like the one in Dataset 1 measure ~30 features per cell with excellent resolution and that capacity is expected to roughly double in the next few years. Spectral flow cytometry is thus well-matched to studies of very low frequency cells, as was the case in Dataset 1, where a goal was to computationally pinpoint hundreds of virus-specific T cells in datasets of over 5 million collected cells.

Datasets 1 and 2 were from individuals infected with two different respiratory viruses, rhinovirus or SARS-Cov-2, respectively. Respiratory viruses are ubiquitous and while some, like rhinovirus, are generally benign, they nonetheless pose risks to patients with underlying chronic health conditions.

The common colds associated with rhinovirus are characterized by shifts in very rare cells in the blood (Muehling et al., 2020; Muehling et al., 2018). In contrast, novel respiratory viruses, such as SARS-CoV-2, the coronavirus causing COVID-19, continue to emerge that enact high morbidity and mortality, even among healthy subjects. Understanding the immune response to such viruses is vital to treatment and vaccine design, and there has been rapid progress applying human immune monitoring to COVID-19 patients (Mathew et al., 2020a; Mathew et al., 2020b; Rodriguez et al., 2020). An ongoing challenge in the field is to quantitatively compare novel diseases, like COVID-19, to other disease states and immune responses. T cells are pivotal to such responses. Severe COVID-19 has been linked to a pathogenic "cytokine storm" in which cellular immune responses likely play a crucial role (Ragab et al., 2020). Nonetheless, in the case of both rhinovirus and COVID-19, it is clear that host factors are a key determinant of the degree of the T cell response (Mathew et al., 2020b; Muehling et al., 2020). By tracking the $CD4^+$ T cells that expand rapidly during infections and respond to immunotherapy, it may be possible to pinpoint or therapeutically guide cells into helpful vs. harmful roles or niches. Overall, a goal of this study was to develop an automated, quantitative toolkit for immune monitoring that would span a wide range of possible immune changes, identify and phenotype statistically significant cell subsets, and provide an overall vector of change indicating both the direction and magnitude of shifts, either in the immune system as a whole or in a key cell subpopulation.

Results

We report here Tracking Responders Expanding (T-REX), a novel unsupervised machine learning algorithm for characterizing cells in phenotypic regions of significant change in a pair of samples (Figure 1). The primary use case for developing the T-REX algorithm was a new dataset from individuals infected with rhinovirus, where changes in the peripheral immune system are expected in very rare memory cells responding directly to the virus (Dataset 1). Infection with rhinovirus is known to induce expansion of circulating virus-specific CD4+ T cells in the blood, and a key feature of the new rhinovirus dataset here is that rare and mechanistically important virus-specific cells were marked with

MHC II tetramers. The T-REX algorithm was blinded to these key features during analysis so that they could subsequently be used to test algorithm performance. In addition, T-REX was tested with paired samples from patients with moderate or severe COVID-19 (Dataset 2), melanoma patients being treated with α-PD-1 checkpoint inhibitor therapies (Dataset 3), and acute myeloid leukemia patients undergoing induction chemotherapy (Dataset 4). These datasets were used to determine whether the T-REX algorithm functions effectively across a spectrum of human immune monitoring challenges and to see how the algorithm performs when changes are restricted to rare cell subsets, as in Dataset 1 and Dataset 3, or when many cells may be expanding or contracting, as in Dataset 2 and Dataset 4.

T-REX identifies cells in phenotypically distinct regions of significant change

For the rhinovirus challenge study in Dataset 1, sample pairs available for T-REX included cells taken immediately prior to infection (i.e., pre-infection, day 0), as well as those during (day 7) or following infection (day 28). Cells were subsampled equally from each time and then concatenated for a single UMAP specific to each analysis pair. UMAP axes were labeled to indicate they were specific to a comparison for a given individual (Figure 2). Thus, each UMAP comparison was a new run of the algorithm. Although it is also possible to map all sample times or all individuals into a single UMAP for analysis, a key goal here was to imagine a minimal T-REX use case with only a pair of samples from one individual. The features selected for UMAP analysis were intentionally limited to surface proteins in order to test whether suitable features for live cell fluorescence activated cell sorting (FACS) could be identified. Following UMAP, each cell was used as the seed for a KNN search of the local neighborhood within the UMAP axes (i.e., the KNN search was within the learned manifold, as with the analysis in Sconify (Burns et al., 2018) or RAPID (Leelatian et al., 2020)). The k-value for KNN was set to 60 as a starting point based on prior studies and later optimized. For each cell, the KNN region could include cells from either time chosen for analysis, and the percentage of each was calculated to determine the representation of each sampled time in a cellular neighborhood. When cells in regions of expansion (≥95% of cells in the KNN region from one sampling time) were clustered together in one

phenotypic region of the UMAP, they were considered a 'hotspot' of significant change. Cells in change hotspots were aggregated and the phenotype automatically characterized using Marker Enrichment Modeling (MEM) (Diggins et al., 2017). MEM labels here indicated features that were enriched relative to a statistical null control on a scale from 0 (no expression or enrichment) to +10 (greatest enrichment). Ultimately, T-REX and MEM were used to reveal hotspots of ≥95% change and assign a label that could be used by experts to infer cell identity.

In the human rhinovirus challenge study yielding Dataset 1, MHC class II tetramers were used to identify rhinovirus-specific $CD4^+$ T cells with the goal of tracking phenotypic changes over the course of infection. Increases in $tetramer^+$ cells on day 7 (Figure 2A) corresponded to the acute infection phase (Muehling et al., 2018). This tetramer tracking system for virus-specific T cells provided an opportunity to test whether the cells identified by T-REX were biologically significant by leaving the tetramer stain features out of the computational analysis (i.e., not using tetramers to make the UMAP or in other parts of T-REX) and then testing to see whether hotspots of cellular change identified by T-REX were statistically enriched for $tetramer^+$, virus-specific cells. In the example subject shown, the pairwise comparisons used in T-REX analysis included $CD4^+$ T cells from day 0, immediately prior to rhinovirus infection, and day 7, a well-studied time point at which rare, virus-specific $CD4^+$ T cells are observed at higher frequencies (Muehling et al., 2018). This trajectory of virus-specific cell expansion was confirmed by a peak in the $\log_2$ fold change in percentage of $tetramer^+$ $CD4^+$ T cells (Figure 2A). Applying T-REX to the rhinovirus data revealed that KNN regions with expansion from day 0 to day 7 were greatly enriched for $tetramer^+$ cells, as compared to regions with less expansion (Figure 2B). UMAP axes were labeled as UMAP_RV-N001_7_0 to denote this UMAP analyzed day 0 and day 7 for individual RV-N001 (Figure 2C). Regions of contraction were observed but were not enriched for $tetramer^+$ cells in any individuals studied here (Figure 3). Notably, two of the nine study subjects challenged with rhinovirus were not infected (N022 and N004); other individuals were infected (Supplemental Table 1).

A key question for the T-REX algorithm is where to set a statistical cutoff for what is considered to be a biologically significant amount of expansion. Two change cutoffs were tested with subject RV-

N001, ≥90% and ≥95% (Figure 2C). Using a cutoff of ≥95% identified both tetramer+ hotspots for RV-N001 and did not identify any additional regions that were not tetramer hotspots (Figure 2C). Thus, ≥95% represented a stringent cutoff that still captured biologically significant cells. An analysis of tetramer enrichment as a function of percentile of expansion from day 0 to day 7 (Figure 2B) showed that tetramer$^+$ cells were not commonly observed to be in local neighborhoods around cells with change below 80% in their KNN region. In contrast, above 90% change, the median CD4$^+$ T cell had 10% or more tetramer$^+$ neighbors around it in the KNN region (Figure 2B). Thus, only regions of 80% or more expansion from day 0 to day 7 were enriched for tetramer$^+$ CD4$^+$ T cells in study individual RV-N001. The statistical cutoff for "greatly expanded" was set at 95% for future analyses.

Regions of significant change contained rhinovirus-specific CD4$^+$ T cells in Dataset 1

The association between change and virus-specific cells observed in the example subject shown (Figure 2B) was observed in five additional rhinovirus subjects; tetramer$^+$ CD4$^+$ T cells were not enriched in KNN regions around cells that had not expanded from day 0 to day 7 (Supplemental Figure 1). This observation suggested that cutoffs at the 5th and 95th percentile would accurately capture cells representing phenotypic regions with significant change over time. In addition, 15th and 85th percentiles were chosen as cutoffs to capture a more moderate degree of change and track cells that might still be of interest but not from regions experiencing significant change. The remaining cells in phenotypic regions between the 15th and 85th percentiles were not considered to have not changed significantly in the context of these studies.

Going forward, it was of interest to determine how often regions of significant change (i.e., the 95th and 5th percentile cutoffs) would contain tetramer$^+$ CD4$^+$ T cells in different individuals participating in the rhinovirus challenge study. As before, tetramer staining was not used in the T-REX analysis to make the UMAP or any other step and was left aside as a way of asking whether the regions of significant change tended to include biologically relevant cells like virus-specific T cells. Cells in regions of significant expansion (≥95%) were also from regions that were also from regions that were enriched

for virus-specific cells in nearly all RV-infected individuals (70%, N = 5 of 7) (Figure 2, Supplemental Figure 1, Figure 3). Thus, by focusing specifically on cells in regions representing the most change over time, T-REX analysis revealed subpopulations containing virus-specific cells. This highlights the ability of T-REX to pinpoint such cells without the use of antigen-specific reagents. For example, for subject RV-N001, 100% (2/2) of tetramer^{+} hotspots were identified automatically using T-REX (Figure 2C).

To determine whether virus-specific cells were identified with this method, all tetramer^{+} regions were also observed in Figure 2C. In analysis of RV-N001, 66.6% (2/3) of tetramer+ hotspots were captured, meaning there was one region with lower change that contained tetramer+ cells. However, there were only 3 cells in these missed regions, confirming that T-REX captured the majority of virus-specific cells in the data set. Following T-REX, MEM analysis was performed using all available features, including some intracellular features not used to define the UMAP space, such as TCF1. The phenotype of the regions of significant change enriched for virus-specific cells was quantitatively described with MEM scores (hotspot 1: ▲CD45R0^{+9} CD38^{+8} ICOS^{+5} CD27^{+4} TCF1^{+4} CCR7^{+4} CCR5^{+3} CD95^{+3} CXCR3^{+2} PD-1^{+2} CD25^{+2} CXCR5^{+2}; hotspot 2: ▲CD45R0^{+9} CD38^{+7} ICOS^{+5} CCR5^{+4} TCF1^{+4} CD27^{+3} PD-1^{+3} CD95^{+3} CXCR3^{+2} CXCR5^{+2}). The change hotspots thus contained cells with central and effector memory signatures.

In the case of an emerging infectious disease, it may not be possible to have a pre-infection sample and it would be useful to know whether T-REX analysis of change between a peak of infection and a later time might also reveal virus-specific T cells. To test this idea, pairwise comparisons were performed with cells from day 7 following RV inoculation and at day 28 after infection (Figure 4). Strikingly, cells in phenotypic regions of significant change again were enriched for virus-specific, tetramer^{+} CD4^{+} T cells. The MEM values for these cells further identified them as CD45R0^{+} memory cells enriched for CD38, ICOS, CD27, TCF1, CXCR5, PD-1, and CD95 expression, a phenotype matching that of the cells identified in the day 0 to day 7 analysis for this individual (RV-N001, Figure 4).

A *k*-value of 60 effectively identified immune hotspots in T-REX

A critical question for KNN analysis is the value of k, the number of neighbors to assess. While it is useful to have a lower *k*-value as the analysis will complete more quickly, increasing the *k*-value might better represent the phenotypic neighborhood or be more statistically robust. To assess how *k*-value impacted detection of cells in regions of change and the degree to which these cells were virus-specific in rhinovirus challenge Dataset 1, the k-value was systematically changed. In example case RV-N001, an optimal *k* was determined to be the inflection point in a graph of the average tetramer enrichment (y-axis, Figure 5) versus increasing values of *k* (x-axis, Figure 5). To calculate this curve, a KNN search was repeated while increasing *k* in steps from 0 to 300 for every cell in each sampling. This analysis was performed for all tetramer$^+$ cells from day 7 (dark purple, Figure 5), all tetramer$^+$ cells from day 0 (light purple, Figure 5), and, as a negative control, random tetramer$^-$ cells from day 7 (black, Figure 5). Within each of these neighborhoods, tetramer enrichment was calculated. This approach identified the inflection point of the tetramer$^+$ density curve as k = 60 for RV-N001 (Figure 5). This *k*-value was used in all other analyses shown in this study, including for other rhinovirus subjects (Figure 3), as well as for COVID-19 patients in Dataset 2 and melanoma and AML patients from Dataset 3 and Dataset 4, respectively, introduced subsequently.

T-REX tracking of direction and degree of change contextualizes diverse immune responses

To consolidate and compare findings for all disease settings analyzed, metrics for degree of change as well as direction of change in each sample were calculated (Figure 6A). Degree of change was calculated as the sum of the percent of cells in the 5th and 95th hotspots of change. Direction of change was calculated as the difference between the number of cells in the 95th and 5th hotspots of change divided by the sum of the number of cells in the 95th and 5th hotspots of change. This way of looking at the data provided a method for comparing changes in many disease types. Rhinovirus subjects had the smallest changes in samples over time with a median of 0.025% and an interquartile

range (IQR) of 0.046% (on a magnitude of change scale from 0 to 100%) across all datasets analyzed. Rhinovirus also had large directionality across all subjects either up or down, with a median of 0.82 and an IQR of 2.00 on the directionality scale from -1.00 to +1.00). Thus, rhinovirus displayed an extremely low magnitude of change in the overall immune system, as very rare cell subsets were responding, and the direction of this change was typically fairly high or low for a given individual (i.e., the changes were not balanced and tended to represent marked expansion or contraction in the rare subsets that changed; Figure 6).

Regions of change included cells expressing CD147 and CD38 in COVID-19 Dataset 2

To test T-REX on other disease settings, the algorithm was applied to Dataset 2, a mass cytometry study of longitudinal collection of blood from patients with COVID-19 (Rodriguez et al., 2020). This study originally contained data for 39 total patients, of which 12 patients had accessible mass cytometry data with at least two blood samples over time. For each patient, the day 0 timepoint and the closest sampled timepoint to day 7, were used for pairwise comparison using T-REX. The COVID-19 samples varied from <1% to the 68% in terms of degree of change with a median of 6.86% and an IQR of 30.4%. The directionality of change was near zero, with a median of -0.00880 and IQR of 0.773. Thus, the blood of COVID-19 patients could display significant changes or little change. Notably, the changes <5% were generally positive (median of 0.55, N = 6), whereas the COVID-19 patient cell populations experiencing change >5% typically decreased between day 0 to day 7 (median directionality of -0.33, N = 6).

In T-REX analysis looking at change on the UMAP axes, patients with significant change were apparent due to large islands of cells being painted dark red or dark blue, indicating >95% change between paired days (Figure 6). These cell populations were clustered and separated into populations representing day 0 or the later time near day 7 and MEM labels calculated in order to assess the identity and phenotypic changes. For example, patient COV26 saw little change (magnitude of 2.02%) that was almost entirely expansion (directionality of 0.99). The largest population experiencing significant

change from COV26 decreased over time and had a MEM phenotype of $CD147^{+10}$ $CD99^{+8}$ $CD29^{+6}$ $CD38^{+4}$ $CD55^{+3}$ $CD14^{+2}$ $CD39^{+2}$ $CD64^{+1}$ $CD56^{+1}$ $CD8a^{+1}$, indicating it was a $CD14^{+}$ myeloid cell subset with high expression of CD147/Basigin all cell phenotypes listed in Supplemental Table 2. The phenotype for all automatically identified clusters of cells that expanded or contracted greatly and the degree and direction of change for each COVID-19 patient from Dataset 2 is listed in Supplemental Table 2. These reference phenotypes should be comparable to those in other studies of COVID-19, and a meta-analysis of phenotypes could use quantitative analysis of MEM labels to compare these highly expanding and contracting cells (Supplemental Table 2).

T-REX reveals immune cell changes during cancer therapies in Dataset 3 and Dataset 4

T-REX was next tested on two previously published cancer immune monitoring studies representing a wide range of immune system changes, from modest to extensive. Dataset 3 consisted of mass cytometry analysis of peripheral blood mononuclear cells from melanoma patients treated with anti-PD-1 (Greenplate et al., 2019). This well-studied dataset primarily includes melanoma patients whose blood had modest, subtle shifts in PBMC phenotypes over time. However, one patient in the set, patient MB-009, developed myelodysplastic syndrome (MDS) and experienced a great shift in blood immunophenotype in parallel with the emergence of a small population of blasts in PBMCs (Greenplate et al., 2016b). Overall, when analyzed by T-REX, the melanoma samples in Dataset 3 for comparisons of day 21/35 versus day 0 had a small degree of change (median of 0.58% and an IQR 2.34%) with a varying directionality (median of -0.42 and an IQR of 1.46) confirming the subtle shifts in phenotypes as previously indicated. The great shift in peripheral immunophenotypes observed in MB-009 was confirmed with T-REX analysis when comparing the 6 week and 12 week times. Notably, at 6 weeks, the peripheral blast count was still below 5% (Greenplate et al., 2016b), so T-REX detected a substantial change in subsets that were not driven solely by the emergence from the marrow of the MDS blasts.

Dataset 4 was chosen to represent large changes and included peripheral blood from AML patients treated with induction chemotherapy (Ferrell et al., 2016). The compared times for the AML data in Dataset 4 were day 5/8 versus day 0. As expected, the majority of AML patients had a large degree of change in samples (median of 81.0% and an IQR of 75.2%) with little to no directionality to the change (median of -0.00250 and an IQR of 0.0173), meaning that there were massive changes in terms of both expansion and contraction over the course of treatment. MEM labels showed that the cells contracting in responder patients were the AML blasts, whereas the emerging cells were the non-malignant immune cells (Figure 6). AML samples with a degree of change >80% (AML001, AML002, AML004) came from patients with high blast count in the blood and complete response to treatment indicating the complete transformation of the immune environment after treatment. AML007, a patient with no blasts in the blood, had a degree of change of 5.97% over treatment. For AML003, a patient that did not respond to treatment, little change was seen from days 0 to 5 (degree of change = 3.19%) by means of T-REX analysis.

Materials and Methods

Generation of Dataset 1

Dataset 1 was a newly generated dataset of PBMCs collected over the course of infection from healthy volunteers who were experimentally challenged with RV. These data were collected and processed at the University of Virginia. Collection times were defined by established kinetics of memory effector T helper cell responses. The study and sample collection were conducted with informed consent, IRB approval, and in accordance with the Declaration of Helsinki. Cells were stained with antibodies detecting key features of helper, naïve, and memory T cells (CCR6, ICOS, CXCR3, CD27, CCR5, TBET, CD45RA, CD45R0, CD95, CXCR5, TCF1, CCR7), antibodies measuring activation and proliferation (CD25, CD38, CD127, Ki-67, PD-1), and up to three MHCII tetramers. RV peptide/MHCII tetramers were used to label and magnetically enrich tetramer-positive cells (Muehling et al., 2016). Data were collected using a 3-laser Aurora spectral flow cytometry instrument.

Data pre-processing

Before testing and evaluating the modular analysis workflow for rare cells, data preprocessing and QC of the data was done on all samples for all time points, which included spectral unmixing with autofluorescence subtraction, spill-over correction, and applying scales transformation. An arcsinh transformation was applied to the dataset with each channel having a tailored cofactor based on the instrument used to acquire the data as well as to stabilize variance near zero. Manual gating for clean-up of the data was done by an expert to exclude debris, doublets, and dead cells. As helper T cells were of interest for this RV study, the data analyzed was manually gated for $CD3^+$ $CD4^+$ T cells.

T-REX algorithm

A modular data analysis workflow including UMAP, KNN, and MEM was developed in R and scripts for analysis of data in this manuscript are available online (https://github.com/cytolab/T-REX). The dimensionality tool used included UMAP, or Uniform Manifold Approximation and Projection. The default parameter settings for UMAP as found in the uwot package in R were used. Since UMAP analyses were specific to a given individual and pair of samples, UMAP axis were labeled to indicate the individual and comparison being made, as in 'UMAP_RV-N001_07', which indicated a comparison of day 0 and day 7 for individual RV-N001. The KNN search from the Fast Nearest Neighbors (FNN) package was used to find the nearest neighbors for a given cell. For this project, a KNN search was done for every cell using the low dimensional projection of the data as an input for the neighborhood search. The value for *k,* or the number of nearest neighbors, was determined by an optimization of tetramer enrichment within a neighborhood.

MEM analysis of enriched features

Marker Enrichment Modeling from the MEM package (https://github.com/cytolab/mem) was used to characterize feature enrichment in KNN region around each cell. MEM normally requires a

comparison of a population against a reference control, such as a common reference sample (Diggins et al., 2017), all other cells (Diggins et al., 2018; Leelatian et al., 2020), or induced pluripotent stem cells (Greenplate et al., 2019). Here, a statistical reference point intended as a statistical null hypothesis was used as the MEM reference. For this statistical null MEM reference, the magnitude was zero and the IQR was the median IQR of all features chosen for the MEM analysis. Values were mapped from 0 enrichment to a maximum of +10 relative enrichment. The contribution of IQR was zeroed out for populations with a magnitude of 0.

Data availability and transparent analysis scripts

Datasets analyzed in this manuscript are available online, including at FlowRepository (Spidlen et al., 2012). COVID-19 Dataset 2 (Rodriguez et al., 2020) (https://ki.app.box.com/s/sby0jesyu23a65cbgv51vpbzqjdmipr1), melanoma Dataset 3 (Greenplate et al., 2016b; Greenplate et al., 2019) (http://flowrepository.org/id/FR-FCM-ZYDG), and AML Dataset 4 (Ferrell et al., 2016; Greenplate et al., 2019) (http://flowrepository.org/id/FR-FCM-ZZMC) were described and shared online in the associated manuscripts. Rhinovirus Dataset 1 is a newly generated dataset created at the University of Virginia and will be made public on FlowRepository before the conclusion of peer review and after removal of protected health information is confirmed.. During peer review, a private reviewer link to Dataset 1 will be used to allow testing of scripts. Transparent analysis scripts for all four datasets and all presented results are publicly available on the CytoLab Github page for T-REX (https://github.com/cytolab/T-REX) and include open source code and commented Rmarkdown analysis walkthroughs.

Discussion

A signature feature of the immune system is the ability of rare cells to respond to a stimulus by activating and proliferating, leading to rapid expansion of highly specialized cells that may share both a distinct phenotype and a clonal origin. The T-REX algorithm was designed to capture phenotypic regions where significant change was occurring between a pair of samples from one individual. The fact that T-REX was able to identify the phenotype of cells whose regions were greatly enriched for virus-specific T cells in the rhinovirus Dataset 1 (Figure 2, Figure 3, Figure 4) was striking from an analysis point of view and closely matches what would be expected based on the current understanding of immunology. Further testing of this result, including genetic analysis for clonal cells in regions that changed but were not enriched for the specific tetramers used in this study is warranted. For example, in subject RV-N006 there were regions of expansion and contraction that were highlighted by T-REX but which were not enriched for tetramer$^+$ cells (Figure 3). This region may contain a clonal response for which a tetramer was not available, or it may contain another type of CD4$^+$ T cell immune response which may or may not be related to the ongoing rhinovirus infection. Additionally, it will be important to test whether this type of finding holds true for other well-studied viruses where tetramers are available, such as influenza (Turner et al., 2020), and whether these findings extend to MHC class I tetramers and CD8 T cells. It was also striking that in the comparisons of day 7 to either day 0 (Figure 2 and Figure 3) or day 28 (Figure 4), only the expanding cells (red) were in regions that were also tetramer hotspots. However, despite the focus on expansion in the T-REX acronym, contracting cells will likely also be of biological significance in different disease settings (as with AML) or potentially at different disease times.

Another result of interest from rhinovirus Dataset 1 was that the regions of great contraction were not enriched for tetramer$^+$ virus-specific CD4$^+$ T cells (Figure 3, MEM phenotypes in Supplemental Table 2). Given that the time point studied was selected to approximate the peak T cell response, it is possible that regions of contraction containing tetramer+ CD4+ T cells would be identified during different infection phases, such as acute cell sequestration and the contraction phase. It is also notable

that in each disease setting, the patient served as an effective baseline for comparison and allowed T-REX to find phenotypically similar cells in individuals with different starting immune profiles (Figure 3). Finally, it is notable that at the times studied here, the virus-specific $CD4^+$ T cells largely bore memory phenotypes that suggested organ trafficking (e.g., CXCR5 enrichment in the MEM labels) and memory (e.g., enrichment of CD45R0 and CD95), as seen in Figure 2, Figure 3, and Figure 4. Comparative phenotyping across time was beyond the scope of this study, but will be of high interest in the future to determine whether the vector of change in specific subsets over time correlates with additional aspects of disease or complicating host factors, such as allergy and asthma.

A central question in systems immune monitoring is to place newly emerging diseases into the context of other well-studied diseases and immune responses. In working to compare COVID-19 and rhinovirus, it became clear that a summary of change indicating both the direction and magnitude of shifts, was needed (Figure 6). This framework represents a way to summarize both broad populations of immune cells, like all $CD45^+$ leukocytes, and key cell subpopulations, like $CD4^+$ T cells. The striking changes observed in patients with moderate and severe COVID-19 were far beyond the subtle changes observed in individuals with rhinovirus and more closely matched the immune reprogramming observed in melanoma patients receiving checkpoint inhibitor therapy (Figure 6). A primary finding of T-REX analysis of Dataset 2 from COVID-19 patient blood was that some patients experienced very large changes in the blood, and that these changes were typically associated with more decreases than increases (Figure 6). This finding closely matches reported findings from others who observed a systematic reprogramming of immune cell populations in many immune cell populations in severe COVID-19 patients (Mathew et al., 2020b). Also observed were T cell subsets with enrichment of CD38, PD-1, and CD95, as has also been previously reported. While disease severity is not available for individual patients from Dataset 2, it is known that all these cases were at least moderate or severe (Rodriguez et al., 2020). It will be of interest to test the hypothesis that the more severe cases will be one of the two groups, either the patients with very little change and just expansion of cells or those with more major change and a general decrease of cells (Figure 6).

Notably, CD147/Basigin, was highly expressed on many cells changing during infection and was observed to change greatly on some populations over time. CD147 has been proposed in pre-prints as both a binding partner for SARS-CoV-2 spike protein and a potential mechanism of cellular entry, although evidence is needed to support this controversial hypothesis (Shilts and Wright, 2020). In the study of Dataset 2, the authors noted that immune responses were dominated by cells expressing CD38 and CD147 (Rodriguez et al., 2020). In the T-REX analysis of the same Dataset 2, for the cells that were changing greatly, CD147 was sometimes present on cells from day 0 that decreased greatly and was lower or absent on cells that emerged only at later times (Figure 6). An example of this was seen in cells from patient COV40, for which the authors noted CD147 expression on effector subsets at 1 week and onwards. The cells pinpointed by T-REX as emerging at day 6 included B cells that expressed CD147 (e.g., CX3CR1^{+8} CD9^{+8} CD29^{+8} CD147^{+5} IgD^{+3} CD99^{+3} CD33^{+1} CD11c^{+1} HLA-DR^{+1} CD24^{+1}, Supplemental Table 2), but the level of enrichment was lower than on myeloid cells from day 0 that decreased over time (e.g., CD147^{+8} CD29^{+6} CD55^{+5} CD38^{+5} CD99^{+4} CD64^{+3} CD62L^{+2} CD45^{+1} CD33^{+1} CD14^{+1}, Supplemental Table 2). This pattern of decreased enrichment of CD147 on cells emerging after day 0 was seen on other patients (Supplemental Table 2) and is consistent with multiple explanations. Overall, there was a strong downward trend in many of the markers and cell subsets in COVID-19 patients, suggesting either selection against cells expressing a high level of proteins, downregulation of expression of key surface markers like CD147, expansion of immature or abnormal cells, or extreme trafficking of cells into tissues. These potential outcomes cannot be distinguished from each other with the analysis here and the utility of the T-REX algorithm is primarily in generating these hypotheses automatically and in pinpointing cells with extreme behavior within the context of the patient as their own baseline. Given the large amounts of change (Figure 6) and the generally lower numbers of T cell subsets observed in COVID-19 than in healthy individuals (Supplemental Table 2), it may be the case that therapeutic stabilization of the immune system will be needed before virus-specific T cells will be identifiable with the T-REX method. It will be especially interesting to explore more mild cases of COVID-19 with this approach and determine whether the hotspots of change are truly virus-

specific.

For the melanoma and AML cases presented here, the cohort sizes were not large enough to allow robust statistical comparison of patient response to degree or direction of change, although this information is available in the original studies (Ferrell et al., 2016; Greenplate et al., 2019). Of the AML patients, those with a high magnitude of change (Figure 6) were also those that had a high blast count and were complete responders to induction therapy, suggesting that the change represents the overall "reset" of the immune system following chemotherapy. It will be of high interest to ask whether the identification of virus-specific T cells extends to populations of cells on checkpoint inhibitor therapy. The dynamics of regulatory cells may also be of interest, especially for autoimmunity, and it is possible but not known whether these cells will follow the same pattern as the $CD4^+$ T cells in rhinovirus.

Once identified, the key features highlighted by T-REX and MEM on the cells in regions undergoing significant change can be used in lower parameter flow cytometry or imaging panels to provide further information, confirm findings, and physically isolate cells by FACS. Thus, low parameter cytometry approaches may rely more on manual analysis methods and cell signatures that are determined *a priori*, and T-REX may provide a useful tool for narrowing in on such features using exploratory high dimensional data. The computational approach here emphasizes unsupervised UMAP and KNN clustering and uses statistical cutoffs to guide the analysis. Further optimization of the algorithm could include a stability testing analysis where the stochastic components of the algorithm are repeated to determine whether clusters or phenotypes are stable (Leelatian et al., 2020; Melchiotti et al., 2017). Overall, the unsupervised approach aims to diminish investigator bias and reveal novel or unexpected cell types. While unsupervised analysis tools have impacted high dimensional cytometry for at least a decade (Amir el et al., 2013; Becher et al., 2014; Bendall et al., 2011; Diggins et al., 2015; Saeys et al., 2016), T-REX is designed to capture both very rare and very common cell types and place them into a common context of immune change. The extremely rare T cells identified here are overlooked by other tools due to these tools typically needing clusters of cells representing at least 1% and generally more than 5% of the sample.

Overall, the idea of mapping individuals with vastly contrasting immune system changes onto a common plot of change (Figure 6) has appeal for placing newly emerging diseases like COVID-19 into the context of existing diseases. This level of sensitivity was striking as it indicated that the T-REX focus on ≥95% expansion had revealed extremely rare, virus-specific CD4+ T cells without prior knowledge of their phenotype. While it is the case that there are different reasons individuals will experience significant change in their immune system – as with the AML patients responding to chemotherapy and the COVID-19 patients with severe disease both have significant change over a short time – this framework goes beyond prior systems immunology approaches, such as earth mover's distance, which provide a non-directional degree of change (Greenplate et al., 2019; Orlova et al., 2016). Thus, T-REX and the resulting direction and magnitude plots represent a useful minimum of information that can be the nidus for further development of systems immunology tools for characterizing responses in rare cells and bulk systems.

Acknowledgements

We thank Monika Grabowska at Vanderbilt University for helpful work during a rotation on code that became part of the T-REX algorithm. Research was supported by the following funding resources: U01 AI125056 (S.M.B., A.G.A.P, L.M.M., J.A.W., and J.M.I.), R01 CA226833 (J.M.I., S.M.B.), U54 CA217450 (J.M.I.), T32 AI007496 (L.M.M.) and the Vanderbilt-Ingram Cancer Center (VICC, P30 CA68485).

Author Contributions

S.M.B., A.G.A.P, L.M.M., J.A.W., and J.M.I. designed the data science study. A.G.A.P, L.M.M., J.A.L., W.W.K., R.B.T., and J.A.W. designed human rhinovirus challenge studies. A.G.A.P, L.M.M., and J.A.L. performed experimental work. S.M.B., A.G.A.P, L.M.M., J.A.L., R.B.T., J.A.W., and J.M.I. performed data analysis, developed figures, and wrote the manuscript. S.M.B., A.G.A.P, and L.M.M. compiled patient data. S.M.B and J.M.I. developed R scripts for data analysis and visualization. J.A.W. and J.M.I. provided financial support. All authors contributed in reviewing the manuscript.

Declaration of interests

J.M.I. was a co-founder and a board member of Cytobank Inc. and received unrelated research support from Incyte Corp, Janssen, and Pharmacyclics. A.G.A.P. became an employee. and J.A.L. became a paid consultant of Cytek Biosciences, Inc. after performing this research at University of Virginia. All other authors declare no competing interests.

References

Amir el, A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature biotechnology *31*, 545-552.

Becher, B., Schlitzer, A., Chen, J., Mair, F., Sumatoh, H. R., Teng, K. W., Low, D., Ruedl, C., Riccardi-Castagnoli, P., Poidinger, M.*, et al.* (2014). High-dimensional analysis of the murine myeloid cell system. Nature immunology *15*, 1181-1189.

Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology.

Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., and Snyder-Cappione, J. E. (2019). Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature communications *10*, 5415.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir el, A. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I.*, et al.* (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science *332*, 687-696.

Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences of the United States of America *111*, E2770-2777.

Burns, T. J., Nolan, G. P., and Samusik, N. (2018). Continuous visualization of differences between biological conditions in single-cell data. BioRxiv, 337485.

Davis, M. M., Tato, C. M., and Furman, D. (2017). Systems immunology: just getting started. Nat Immunol *18*, 725-732.

Diggins, K. E., Ferrell, P. B., Jr., and Irish, J. M. (2015). Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. Methods *82*, 55-63.

Diggins, K. E., Gandelman, J. S., Roe, C. E., and Irish, J. M. (2018). Generating Quantitative Cell Identity Labels with Marker Enrichment Modeling (MEM). Current protocols in cytometry / editorial board, J Paul Robinson, managing editor [et al] *83*, 10 21 11-10 21 28.

Diggins, K. E., Greenplate, A. R., Leelatian, N., Wogsland, C. E., and Irish, J. M. (2017). Characterizing cell subsets using marker enrichment modeling. Nature methods *14*, 275-278.

Ferrell, P. B., Jr., Diggins, K. E., Polikowsky, H. G., Mohan, S. R., Seegmiller, A. C., and Irish, J. M. (2016). High-Dimensional Analysis of Acute Myeloid Leukemia Reveals Phenotypic Changes in Persistent Cells during Induction Therapy. PloS one *11*, e0153207.

Ferrer-Font, L., Mayer, J. U., Old, S., Hermans, I. F., Irish, J., and Price, K. M. (2020). High-Dimensional Data Analysis Algorithms Yield Comparable Results for Mass Cytometry and Spectral Flow Cytometry Data. Cytometry Part A : the journal of the International Society for Analytical Cytology.

Gandelman, J. S., Byrne, M. T., Mistry, A. M., Polikowsky, H. G., Diggins, K. E., Chen, H., Lee, S. J., Arora, M., Cutler, C., Flowers, M.*, et al.* (2019). Machine learning reveals chronic graft-versus-host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies. Haematologica *104*, 189-196.

Greenplate, A. R., Johnson, D. B., Ferrell, P. B., Jr., and Irish, J. M. (2016a). Systems immune monitoring in cancer therapy. European journal of cancer *61*, 77-84.

Greenplate, A. R., Johnson, D. B., Roussel, M., Savona, M. R., Sosman, J. A., Puzanov, I., Ferrell, P. B., and Irish, J. M. (2016b). Myelodysplastic Syndrome Revealed by Systems Immunology in a Melanoma Patient Undergoing Anti-PD-1 Therapy. Cancer Immunol Res *4*, 474-480.

Greenplate, A. R., McClanahan, D. D., Oberholtzer, B. K., Doxie, D. B., Roe, C. E., Diggins, K. E., Leelatian, N., Rasmussen, M. L., Kelley, M. C., Gama, V.*, et al.* (2019). Computational Immune Monitoring Reveals Abnormal Double-Negative T Cells Present across Human Tumor Types. Cancer Immunol Res *7*, 86-99.

Irish, J. M. (2014). Beyond the age of cellular discovery. Nature immunology *15*, 1095-1097.

Irish, J. M., Kotecha, N., and Nolan, G. P. (2006). Mapping normal and cancer cell signalling networks: towards single-cell proteomics. Nature reviews Cancer *6*, 146-155.

Krijthe, J., van der Maaten, R., and Rtsne, L. (2015). L. T-distributed stochastic neighbor embedding using a barnes-hut implementation. In.

Leelatian, N., Sinnaeve, J., Mistry, A. M., Barone, S. M., Brockman, A. A., Diggins, K. E., Greenplate, A. R., Weaver, K. D., Thompson, R. C., Chambless, L. B.*, et al.* (2020). Unsupervised machine learning reveals risk stratifying glioblastoma tumor cells. eLife *9*.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir el, A. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R.*, et al.* (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184-197.

Mathew, D., Giles, J. R., Baxter, A. E., Greenplate, A. R., Wu, J. E., Alanio, C., Oldridge, D. A., Kuri-Cervantes, L., Pampena, M. B., D'Andrea, K.*, et al.* (2020a). Deep immune profiling of COVID-19 patients reveals patient heterogeneity and distinct immunotypes with implications for therapeutic interventions. bioRxiv.

Mathew, D., Giles, J. R., Baxter, A. E., Oldridge, D. A., Greenplate, A. R., Wu, J. E., Alanio, C., Kuri-Cervantes, L., Pampena, M. B., D'Andrea, K.*, et al.* (2020b). Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426.

Melchiotti, R., Gracio, F., Kordasti, S., Todd, A. K., and de Rinaldis, E. (2017). Cluster stability in the analysis of mass cytometry data. Cytometry Part A : the journal of the International Society for Analytical Cytology *91*, 73-84.

Mistry, A. M., Greenplate, A. R., Ihrie, R. A., and Irish, J. M. (2019). Beyond the message: advantages of snapshot proteomics with single-cell mass cytometry in solid tumors. The FEBS journal *286*, 1523-1539.

Muehling, L. M., Heymann, P. W., Wright, P. W., Eccles, J. D., Agrawal, R., Carper, H. T., Murphy, D. D., Workman, L. J., Word, C. R., Ratcliffe, S. J.*, et al.* (2020). Human TH1 and TH2 cells targeting rhinovirus and allergen coordinately promote allergic asthma. The Journal of allergy and clinical immunology.

Muehling, L. M., Mai, D. T., Kwok, W. W., Heymann, P. W., Pomés, A., and Woodfolk, J. A. (2016). Circulating Memory CD4⁺ T Cells Target Conserved Epitopes of Rhinovirus Capsid Proteins and Respond Rapidly to Experimental Infection in Humans. The Journal of Immunology, 1600663.

Muehling, L. M., Turner, R. B., Brown, K. B., Wright, P. W., Patrie, J. T., Lahtinen, S. J., Lehtinen, M. J., Kwok, W. W., and Woodfolk, J. A. (2018). Single-Cell Tracking Reveals a Role for Pre-Existing CCR5+ Memory Th1 Cells in the Control of Rhinovirus-A39 After Experimental Challenge in Humans. The Journal of infectious diseases *217*, 381-392.

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S.*, et al.* (2016). Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. PloS one *11*, e0151859.

Pyne, S., Lee, S. X., Wang, K., Irish, J., Tamayo, P., Nazaire, M. D., Duong, T., Ng, S. K., Hafler, D., Levy, R.*, et al.* (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. PloS one *9*, e100334.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. In Nat Biotechnol, pp. 886-891.

Ragab, D., Salah Eldin, H., Taeimah, M., Khattab, R., and Salem, R. (2020). The COVID-19 Cytokine Storm; What We Know So Far. Frontiers in immunology *11*, 1446.

Rodriguez, L., Pekkarinen, P., Tadepally, L. K., Tan, Z., Rosat Consiglio, C., Pou, C., Chen, Y., Habimana Mugabo, C., Nguyen Quoc, A., Nowlan, K.*, et al.* (2020). Systems-level immunomonitoring from acute to recovery phase of severe COVID-19. medRxiv, 2020.2006.2003.20121582.

Saeys, Y., Gassen, S. V., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature reviews Immunology *16*, 449-462.

Shilts, J., and Wright, G. J. (2020). No evidence for basigin/CD147 as a direct SARS-CoV-2 spike binding receptor. bioRxiv, 2020.2007.2025.221036.

Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N., and Brinkman, R. R. (2012). FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. Cytometry Part A : the journal of the International Society for Analytical Cytology *81*, 727-731.

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

Turner, J. S., Lei, T., Schmitz, A. J., Day, A., Choreno-Parra, J. A., Jimenez-Alvarez, L., Cruz-Lagunas, A., House, S. L., Zuniga, J., Ellebedy, A. H., and Mudd, P. A. (2020). Impaired Cellular Immune Responses During the First Week of Severe Acute Influenza Infection. The Journal of infectious diseases.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry Part A : the journal of the International Society for Analytical Cytology.

Weber, L. M., and Robinson, M. D. (2016). Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry Part A : the journal of the International Society for Analytical Cytology *89*, 1084-1096.

Figure legends

Figure 1 - Tracking Responders EXpanding (T-REX) algorithm identifies rare cells based on significant expansion or contraction during infection or treatment. Graphic of the Tracking Responders Expanding (T-REX) workflow. Data from paired samples of blood from a subject are collected over the course of infection and analyzed by high dimensional, high cellularity cytometry approaches (e.g., Aurora or CyTOF instrument, as with datasets here). Cells from the sample pair are then equally subsampled for UMAP analysis. A KNN search is then performed within the UMAP manifold for every cell. For every cell, the percent change between the sample pairs is calculated for the cells within its KNN region. Regions of marked expansion or contraction during infection are then analyzed to identify cell types and key features using MEM. For some datasets, additional information not used in the analysis could be assessed to determine whether identified cells were virus-specific. Finally, the average direction and magnitude of change for cells in the sample was calculated as an overall summary of how the analyzed cells changed between samples.

Figure 2 – T-REX identifies molecular signatures of CD4+ T cells that are expanded during acute rhinovirus infection and enriched for virus-specific cells. A subject (RV-N001) was experimentally infected with rhinovirus (RV-A16) and CD4+ T cell signatures monitored by spectral flow cytometry in conjunction with tetramer staining during the course of infection. A) Fold change in the number of tetramer-positive cells (log2) after rhinovirus challenge on day 0. B) Data showing the percentage of tetramer+ cells in each cell's KNN region (where k = 60) plotted against the percentage change in its KNN region on day 7 vs. day 0. A statistical threshold of 90% or higher for the percentage change in KNN region corresponded to marked enrichment of tetramer+ cells at day 7. C) UMAP plots with T-REX analysis of CD4+ T cells for day 7 vs. day 0 based on statistical thresholds of 90-95% change (left column) and ≥95% change (right column) in cell phenotypes. Pink and red colors denote regions of phenotypic change identified by T-REX. Numbers of tetramer+ cells within the cell's KNN region captured in these areas of phenotypic change are denoted. Tetramer analysis revealed that cells labeled pink contained >5% tetramer+ virus-specific cells in the corresponding KNN region. Red cells denote a KNN region that was not enriched for tetramer+ cells, and purple cells denote a tetramer enriched region not captured by T-REX. Values in black indicate the actual number of tetramer+ cells in each circled hotspot of phenotypic change. MEM labels on the right indicate cell phenotypes of each hotspot.

Figure 3 – Cells in regions of significant change between day 0 and day 7 were typically in tetramer^{+} hotspots. T-REX plots of regions of significant change (blue and red) are shown on UMAP axes for CD4+ T cells from 9 rhinovirus challenge study individuals. Solid pink circles indicate tetramer+ hotspots that also contained cells that were in regions of marked expansion. A dashed pink circle indicates the one region of tetramer+ cells that did not exhibit significant change.

Figure 4 – Infected cell phenotypes can be compared to cells taken after infection to reveal regions of expansion. (a) Box and whisker plot shows KNN regions in terms of expansion during

infection represented by percent change as well as percent of tetramer positive cells for post-infection (day 28) and during infection (day 7). (b) UMAP plots for 95 percent change and 5 percent tetramer cutoffs. Cell count is in black as well as in the upper right of each UMAP plot. MEM labels are given for every hotspot of KNN highly expanded and tetramer enriched regions.

Figure 5 – KNN analysis around tetramer$^+$ cells reveals an optimized *k*-value at the inflection point of the tetramer density curve. (a) Tetramer$^+$ cells from day 7 (dark purple) or from day 0 (light purple) and random tetramer$^-$ cells from day 7 (black) are shown overlaid on a common UMAP plot. The number of cells for each group is shown in the upper left of each plot. (b) Average tetramer enrichment is shown for increasing *k*-values in repeated KNN analysis of the cells. The inflection point of the resulting curve is circled in red at $k = 60$, which was the optimized *k*-value for KNN implemented in T-REX.

Figure 6 – Mapping degree and direction of change for 5th & 95th hotspots reveals disease-specific patterns. (A) Degree of change and direction of change from T-REX analysis in a time point comparison shown for AML (day 5/8 vs day 0), COVID (day 1/3/4/5/6/7 vs day 0), MB (day 21/35 vs day 0), and RV (day 7 vs day 0) samples. (B) Example T-REX plots are shown for each disease type analyzed. Degree of change shown in red and blue with red showing regions of expansion over time compared to the blue representing regions of contraction over time. MEM label given for change hotspots in the left example in each sample type.

Figure 1

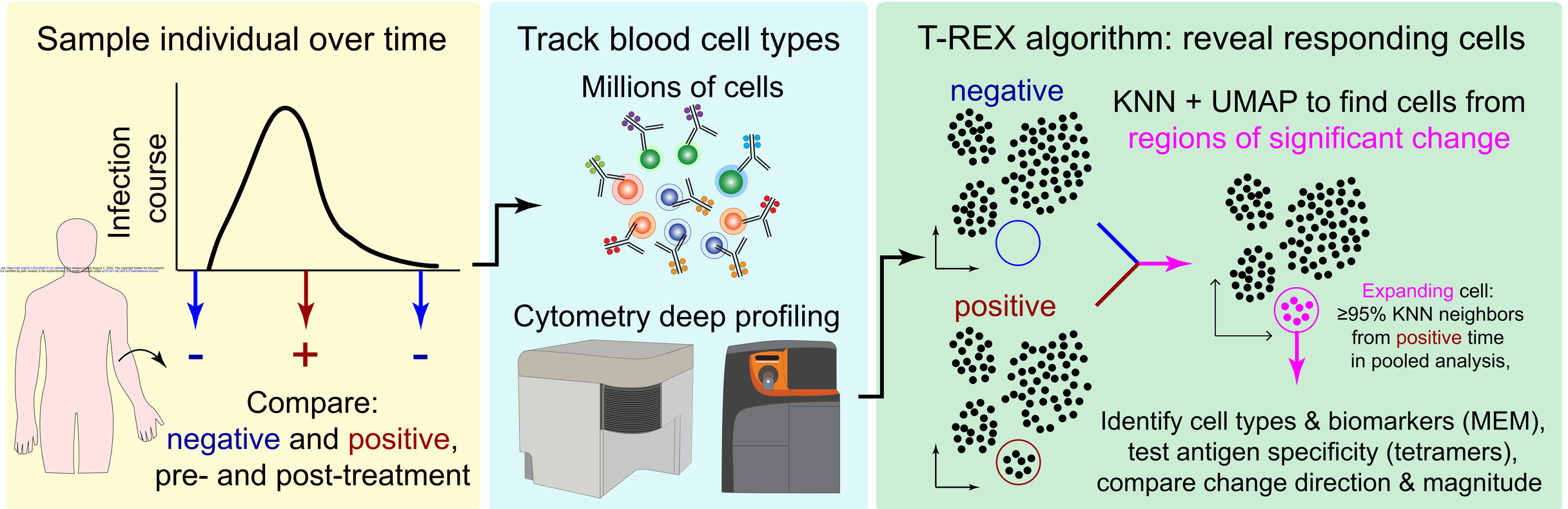
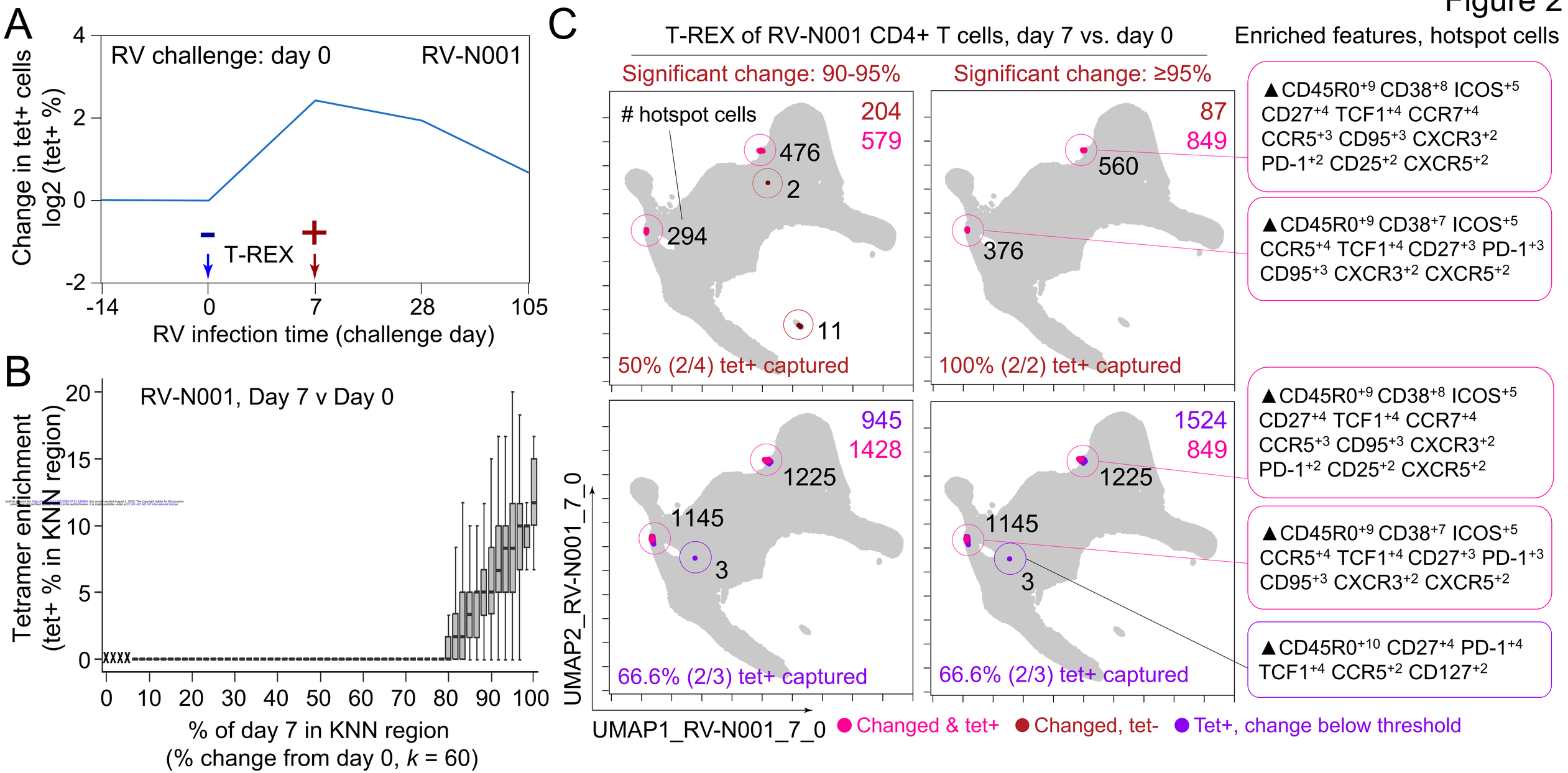
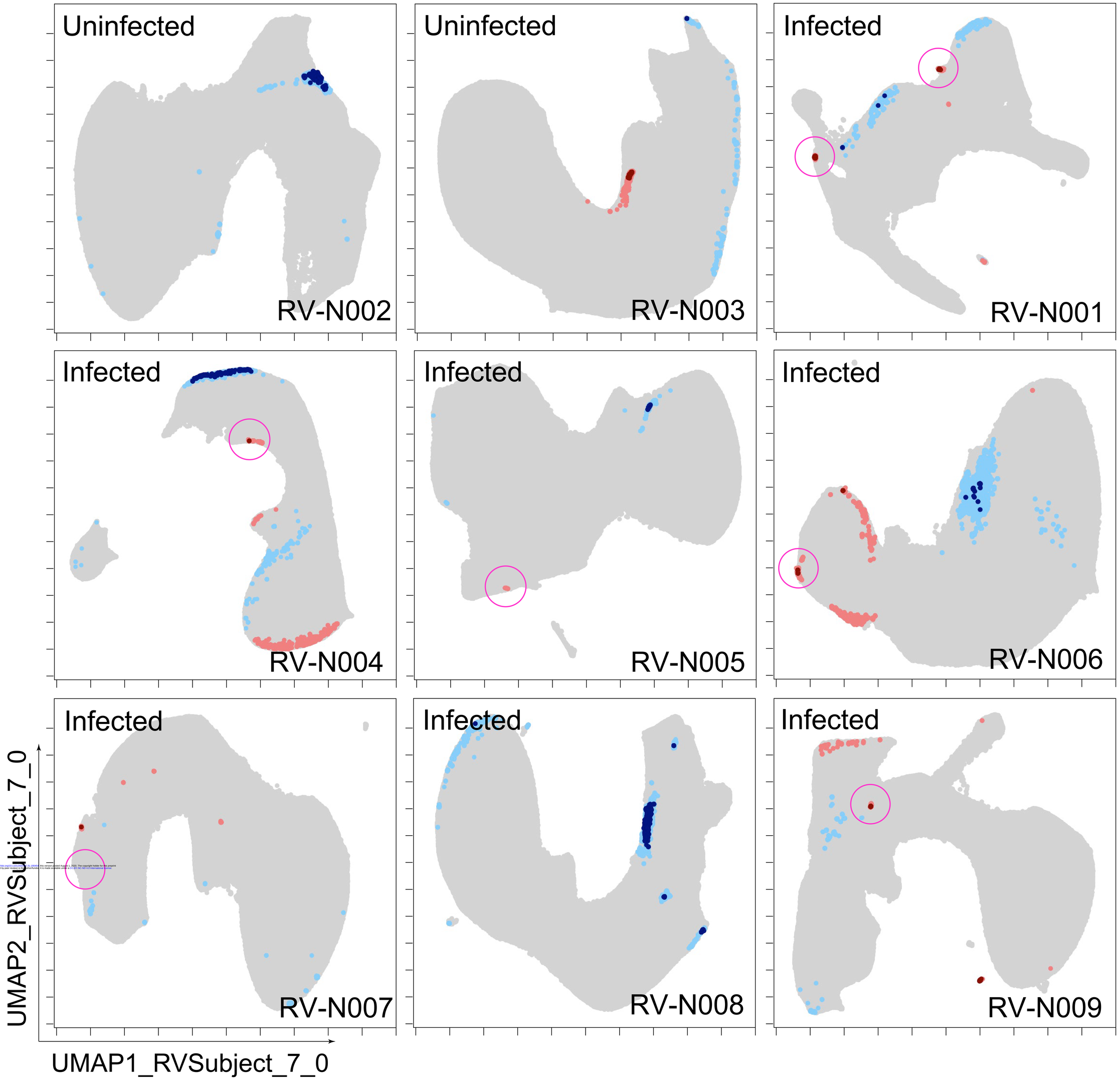


Figure 2

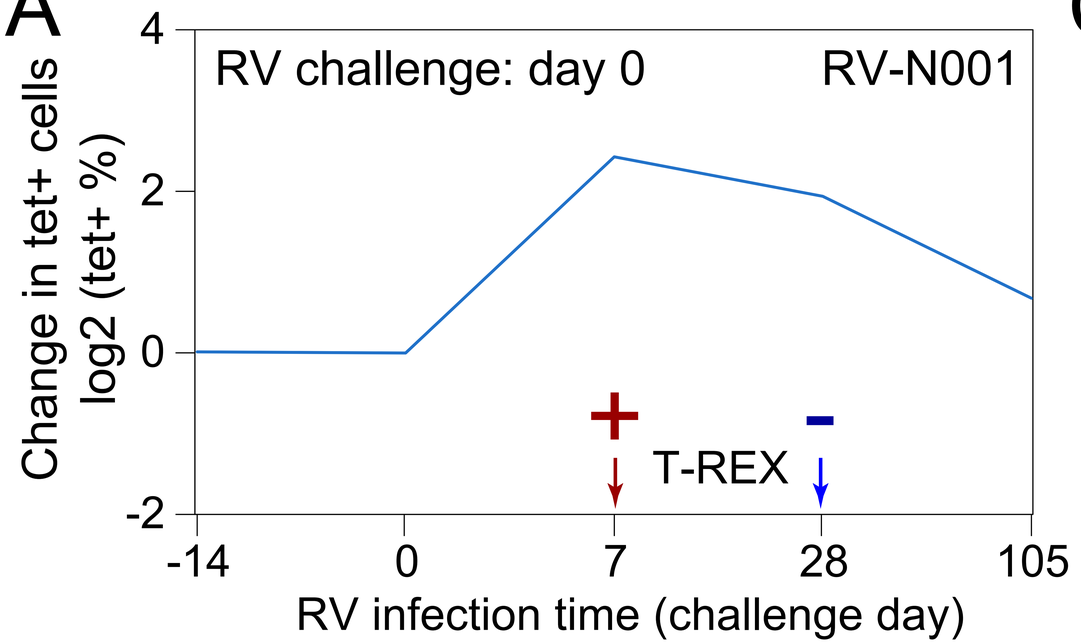

A
RV challenge: day 0
RV-N001
Change in tet+ cells
log2 (tet+ %)
T-REX
RV infection time (challenge day)
B
RV-N001, Day 7 v Day 0
Tetramer enrichment
(tet+ % in KNN region)
% of day 7 in KNN region
(% change from day 0, k = 60)
C
T-REX of RV-N001 CD4+ T cells, day 7 vs. day 0
Significant change: 90-95%
Significant change: ≥95%
hotspot cells
204
579
476
2
294
11
50% (2/4) tet+ captured
87
849
560
376
100% (2/2) tet+ captured
945
1428
1225
1145
3
66.6% (2/3) tet+ captured
1524
849
1225
1145
3
66.6% (2/3) tet+ captured
UMAP2_RV-N001_7_0
UMAP1_RV-N001_7_0
Changed & tet+
Changed, tet-
Tet+, change below threshold
Enriched features, hotspot cells
▲CD45R0+9 CD38+8 ICOS+5 CD27+4 TCF1+4 CCR7+4 CCR5+3 CD95+3 CXCR3+2 PD-1+2 CD25+2 CXCR5+2
▲CD45R0+9 CD38+7 ICOS+5 CCR5+4 TCF1+4 CD27+3 PD-1+3 CD95+3 CXCR3+2 CXCR5+2
▲CD45R0+9 CD38+8 ICOS+5 CD27+4 TCF1+4 CCR7+4 CCR5+3 CD95+3 CXCR3+2 PD-1+2 CD25+2 CXCR5+2
▲CD45R0+9 CD38+7 ICOS+5 CCR5+4 TCF1+4 CD27+3 PD-1+3 CD95+3 CXCR3+2 CXCR5+2
▲CD45R0+10 CD27+4 PD-1+4 TCF1+4 CCR5+2 CD127+2

T-REX analysis of day 7 (expanding, red) and day 0 (contracting, blue) for rhinovirus subjects

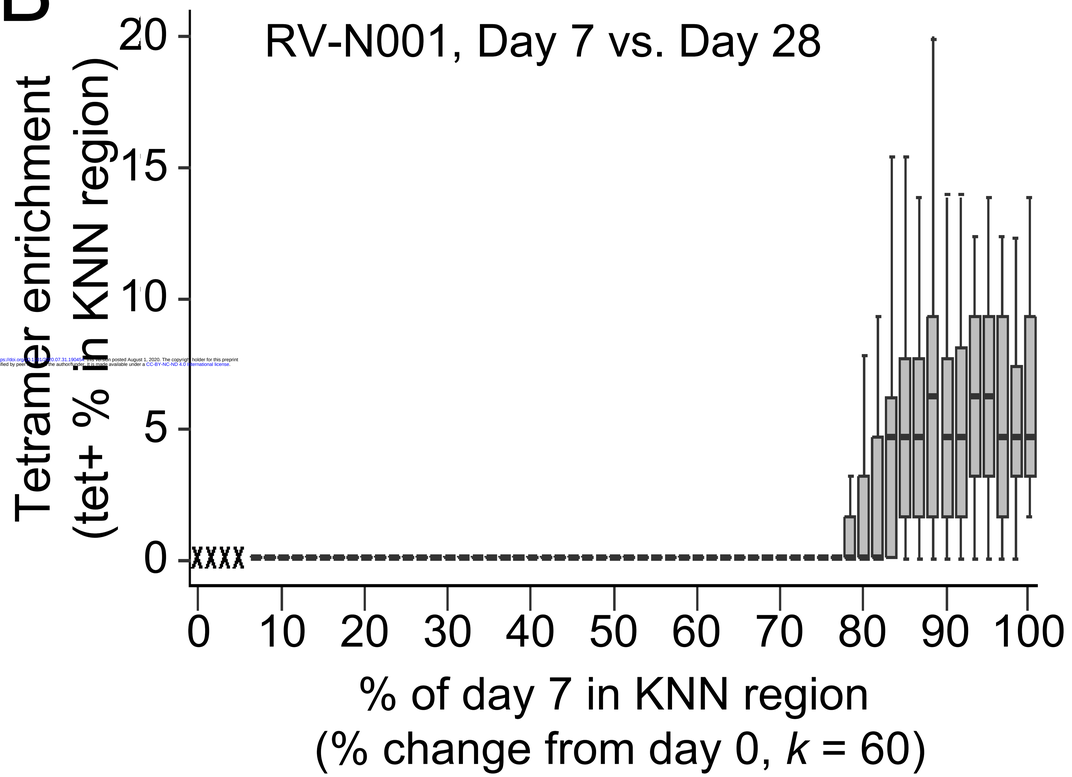
≥95% from day 0 >85% from day 0 >85% from day 7 ≥95% from day 7 Tetramer+ hotspot



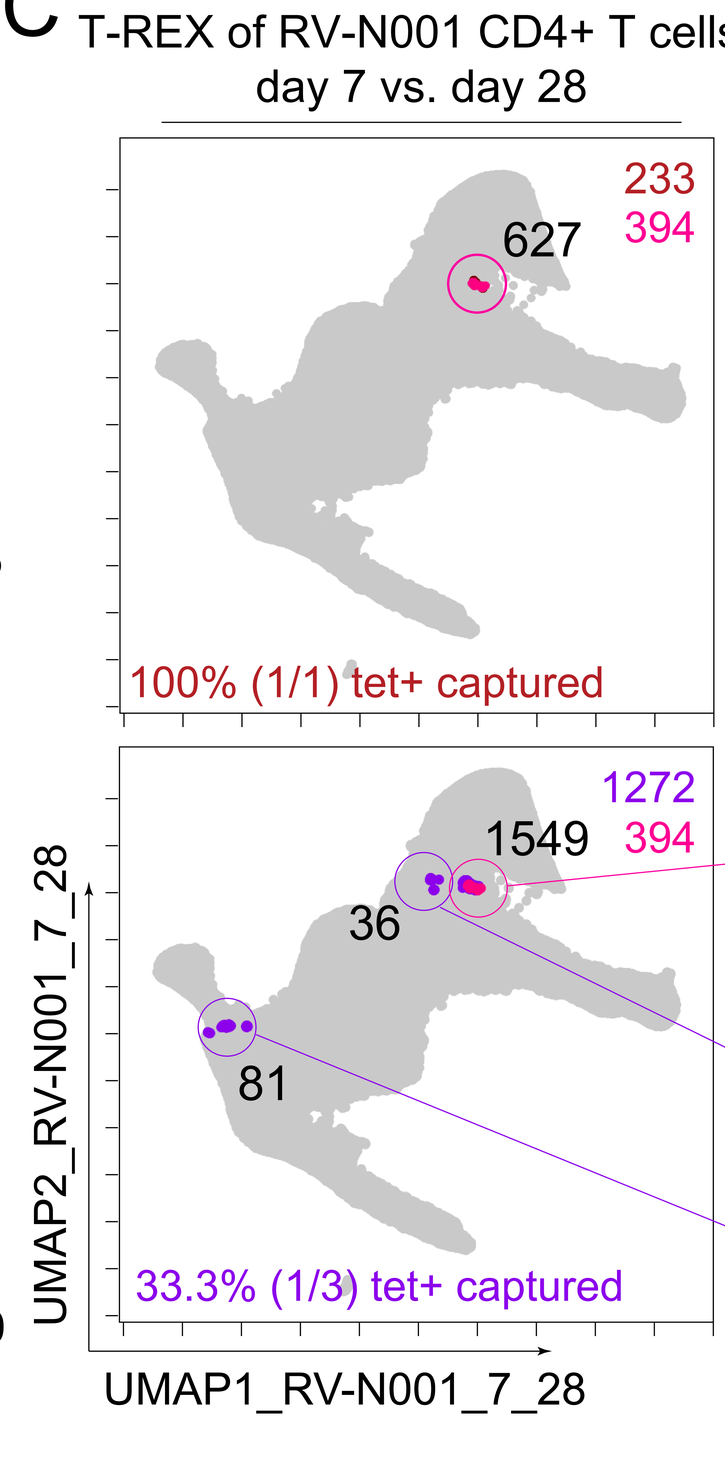
A



B



C



Changed & tet+

Changed, tet-

Tet+, change below threshold

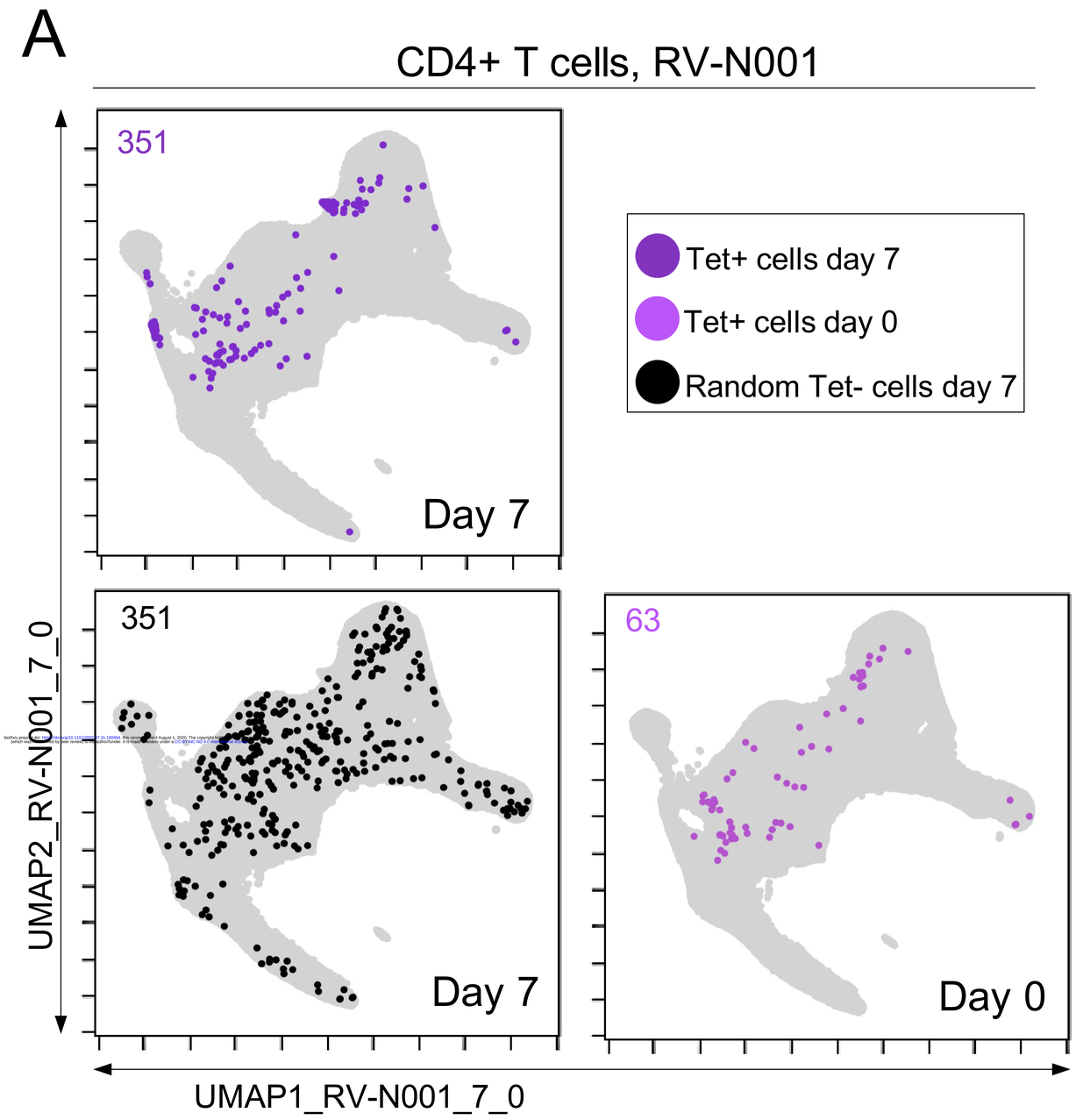
Enriched features, hotspot cells

▲ CD45R0^{+10} CD38^{+7} ICOS^{+5} CD27 $^{+4}$ TCF1^{+4} CCR5^{+3} PD-1^{+3} CD95^{+3} CXCR3^{+2} CXCR5^{+2} CCR7^{+2}

▲ CD45R0^{+10} TCF1^{+5} CD27^{+4} CCR7 $^{+4}$ PD-1^{+3} CD38^{+3} ICOS^{+2}

▲ CD45R0^{+10} TCF1^{+4} CD27^{+3} CD38 $^{+3}$ ICOS^{+2} CCR5^{+2} PD-1^{+2} CD95 $^{+2}$ CCR7^{+2}

Figure 5


A
CD4+ T cells, RV-N001
351
Day 7
Tet+ cells day 7
Tet+ cells day 0
Random Tet- cells day 7
351
Day 7
63
Day 0
UMAP2_RV-N001_7_0
UMAP1_RV-N001_7_0

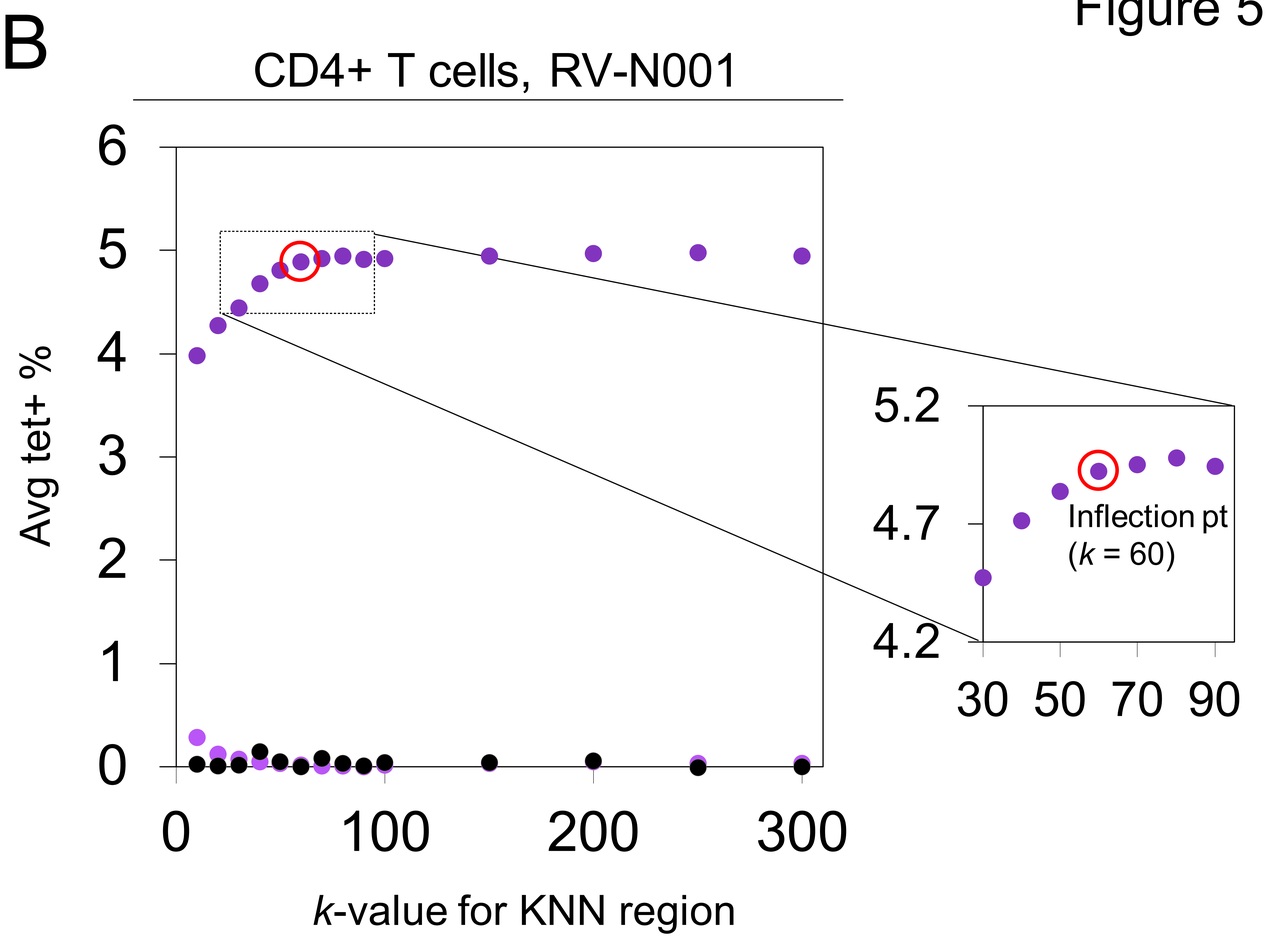
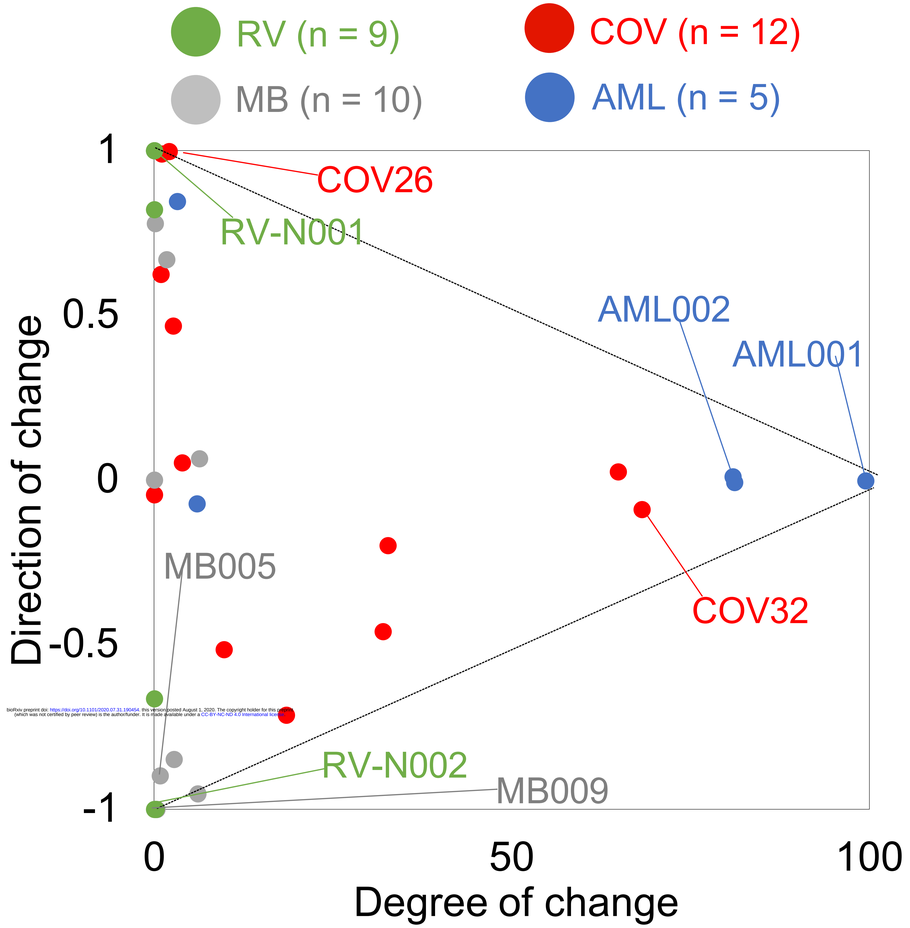
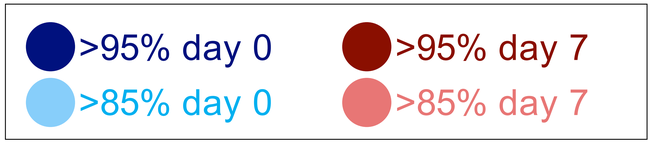

B
CD4+ T cells, RV-N001
Avg tet+ %
0
1
2
3
4
5
6
0
100
200
300
k-value for KNN region
5.2
4.7
4.2
30
50
70
90
Inflection pt
(k = 60)

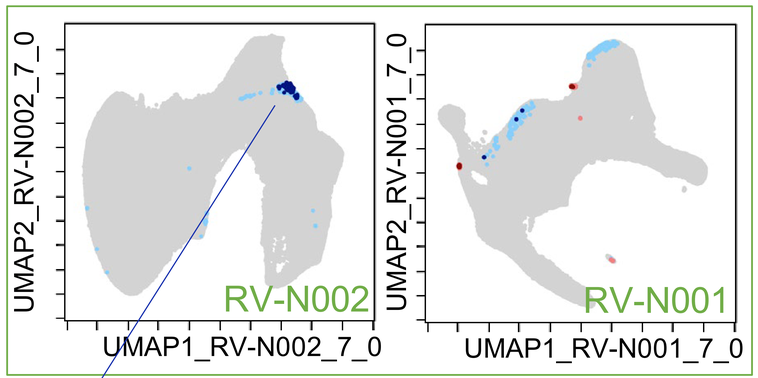
Figure 6

A



B

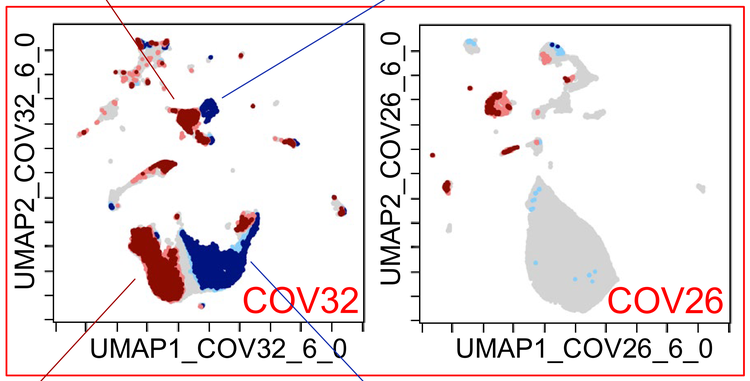




$CD45R0^{+9}$ $CCR7^{+5}$
$CD27^{+4}$ $TCF1^{+4}$
$CD127^{+2}$ $CXCR5^{+2}$

$CD147^{+6}$ $CD29^{+5}$ $CD99^{+5}$
$CD55^{+4}$ $CD11c^{+3}$ $CD9^{+3}$
$CD14^{+3}$ $CD38^{+3}$ $CD39^{+3}$
$CD45^{+2}$ $CD64^{+2}$ $CD56^{+2}$

$CD147^{+10}$ $CD29^{+7}$ $CD38^{+7}$
$CD55^{+6}$ $CD99^{+6}$ $CD64^{+5}$
$CD14^{+5}$ $CD45^{+3}$ $CD11c^{+3}$
$CD62L^{+3}$ $CD56^{+3}$ $CD39^{+3}$

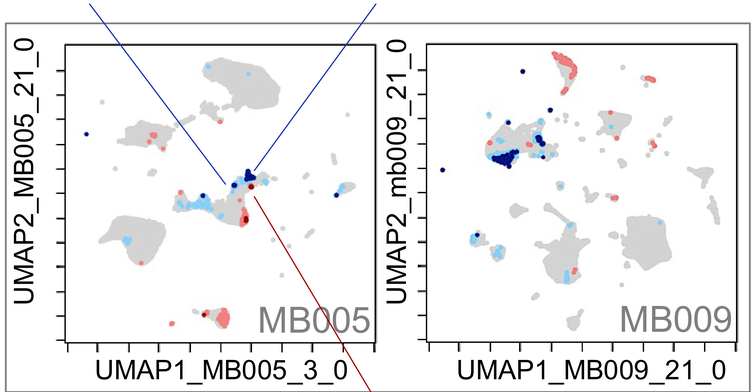


$CD147^{+5}$ $CD24^{+3}$
$CD16^{+3}$ $CD55^{+2}$
$CD29^{+2}$ $CD39^{+2}$

$CD147^{+8}$ $CD55^{+5}$
$CD16^{+5}$ $CD24^{+4}$
$CD29^{+3}$ $CD39^{+3}$

$CD45^{+9}$ $CD43^{+5}$
$CD45R0^{+5}$ $CD44^{+2}$

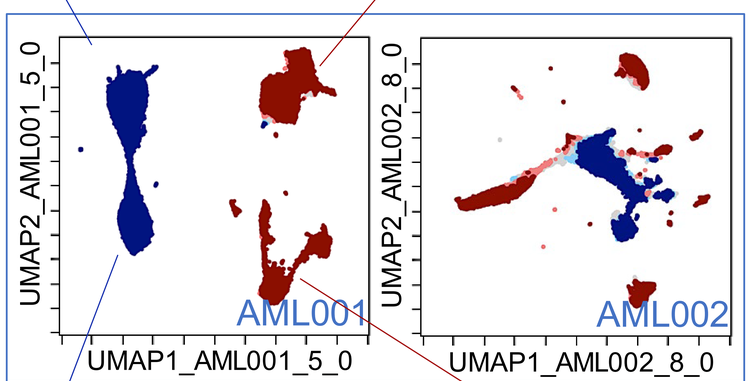
$CD45^{+6}$ $CD16^{+4}$
$CD43^{+2}$ $CD45R0^{+2}$



$CD45^{+7}$ $CD69^{+7}$ $CCR4^{+5}$ $CD43^{+3}$
$CD33^{+3}$ $CD38^{+2}$ $CD45R0^{+2}$

None

$CD3^{+10}$ $CD45^{+6}$ $HLA\text{-}DR^{+3}$
$CD11b^{+2}$ $CD184^{+2}$



$CD45^{+5}$ $CD64^{+4}$ $CD4^{+3}$ $CD33^{+3}$
$CD38^{+3}$ $HLA\text{-}DR^{+3}$ $CD184^{+3}$
$CD11b^{+2}$ $CD123^{+2}$ $CD11c^{+2}$ $CD56^{+2}$

$CD3^{+10}$ $CD45^{+7}$
$CD7^{+3}$ $CD184^{+3}$
$CD4^{+2}$ $CD62L^{+2}$