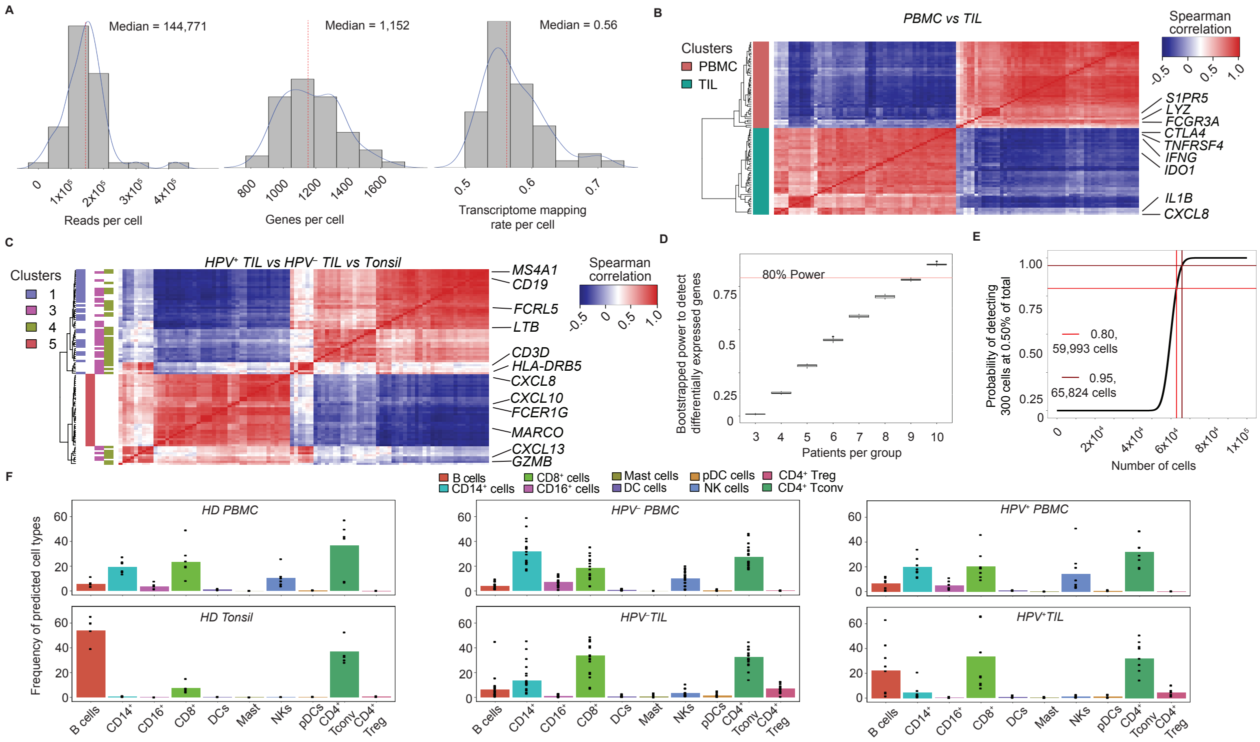


Figure S1. Single-cell RNAseq metrics, post-hoc assessment of power, and frequencies of cells per sample type, related to Figure 1. (A) Single-cell RNAseq metrics for all samples. We sequenced to a median depth of >100,000 reads/cell, recovered a median of >1,000 genes/cell, and had a transcriptome mapping rate of 56%. (B) Pseudobulk RNAseq analysis comparing PBMC versus TIL. Top differentially expressed genes in PBMC were those associated with peripheral blood myeloid cells (e.g. *LYZ* and *FCGR3A*), while those in TIL were reflective of immunoregulatory genes (e.g. *CTLA4* and *IDO1*) and inflammatory cytokines (e.g. *IFNG* and *IL1B*). (C) Pseudobulk RNAseq analysis comparing cluster of tissue resident cells. Clusters 1, 3 and 4 had signatures reflective of B cells (e.g. *MS4A1* and *CD19*) and T cells (*CD3D*), while cluster 5 consisted of genes associated with chemokines (e.g. *CXCL8* and *CXCL10*) and TIL myeloid cells (e.g. *FCER1A* and *MARCO*). (D-E) We performed two post-hoc power analyses to determine the minimum number of samples. First, we performed a bootstrapped power analysis on the pseudobulk data (D), using the mean and standard deviation of each gene across samples to generate synthetic data, and then comparing the number of differentially expressed genes between groups. This analysis revealed that in a balanced study design, we would require 9 samples each from tonsil, HPV$^-$ TIL and HPV$^+$ TIL to recover ~80% of the differentially expressed genes in our study. Second, we used a negative binomial distribution to determine the number of cells required from TIL to have an 80% and 95% probability of detecting 300 cells present at a frequency of 0.5% (which was approximately the frequency of mast cells, i.e. the rarest cell type recovered, in TIL samples). We found that ~60,000 cells would provide an 80% probability of recovering this population of mast cells, and notably our dataset contained 59,599 TIL cells (E). (F) Frequencies of cells identified by single-cell RNAseq from each patient, across all sample groups

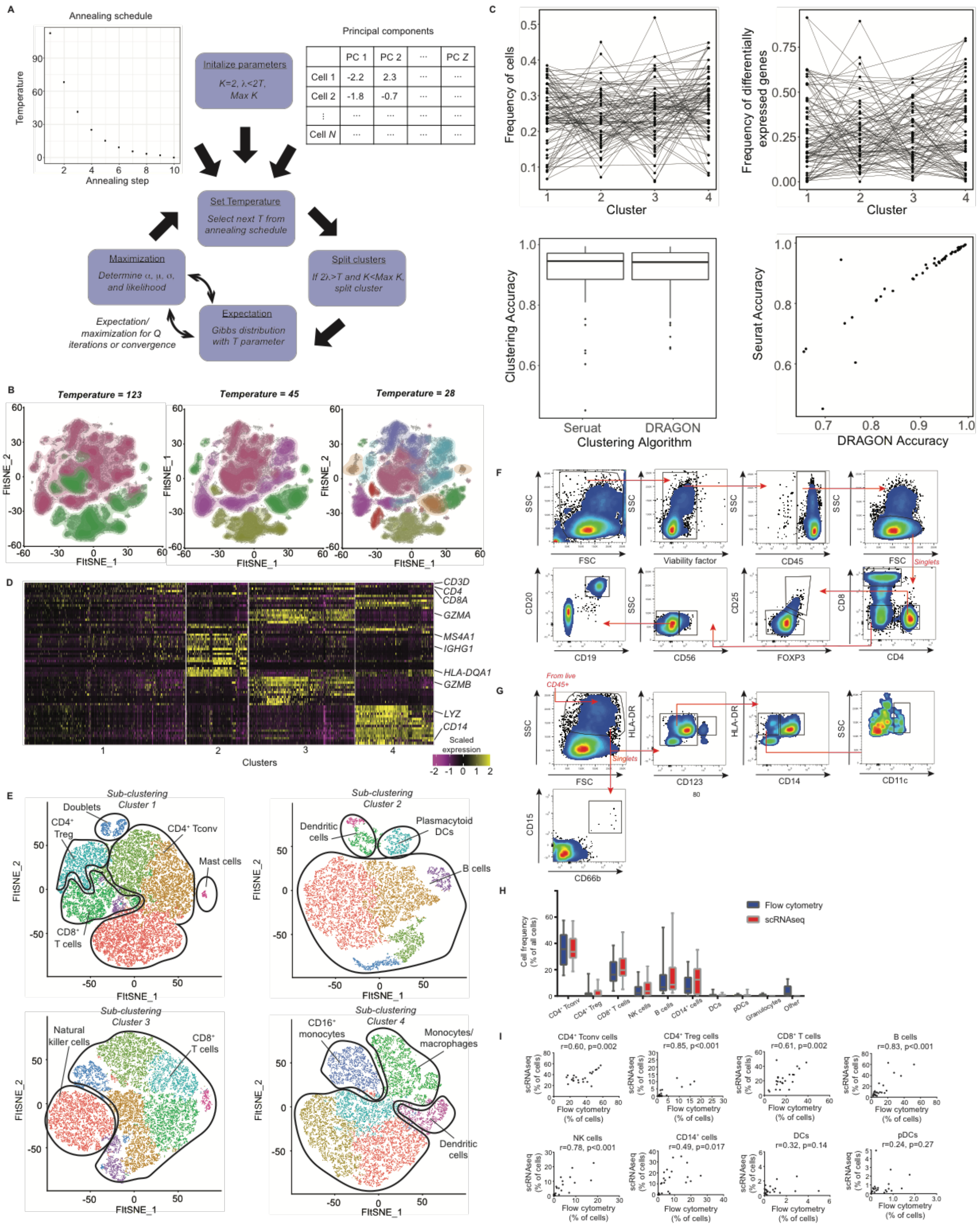

A
Annealing schedule
Temperature
Annealing step
Principal components
PC 1
PC 2
PC Z
Cell 1
Cell 2
Cell N
Initalize parameters
K=2, λ<2T, Max K
Set Temperature
Select next T from annealing schedule
Split clusters
If 2λ>T and K<Max K, split cluster
Expectation
Gibbs distribution with T parameter
Maximization
Determine α, μ, σ, and likelihood
Expectation/ maximization for Q iterations or convergence
B
Temperature = 123
Temperature = 45
Temperature = 28
FltSNE_1
FltSNE_2
C
Frequency of cells
Frequency of differentially expressed genes
Cluster
Clustering Accuracy
Seruat
DRAGON
Clustering Algorithm
Seurat Accuracy
DRAGON Accuracy
D
CD3D
CD4
CD8A
GZMA
MS4A1
IGHG1
HLA-DQA1
GZMB
LYZ
CD14
Scaled expression
Clusters
E
Sub-clustering Cluster 1
Doublets
CD4+ Treg
CD4+ Tconv
Mast cells
CD8+ T cells
Sub-clustering Cluster 2
Dendritic cells
Plasmacytoid DCs
B cells
Sub-clustering Cluster 3
Natural killer cells
CD8+ T cells
Sub-clustering Cluster 4
CD16+ monocytes
Monocytes/ macrophages
Dendritic cells
F
SSC
FSC
Viability factor
CD45
Singlets
CD20
CD19
CD56
CD25
FOXP3
CD8
CD4
G
From live CD45+
HLA-DR
CD123
CD14
CD11c
CD15
CD66b
H
Cell frequency (% of all cells)
Flow cytometry
scRNAseq
CD4+ Tconv
CD4+ Treg
CD8+ T cells
NK cells
B cells
CD14+ cells
DCs
pDCs
Granulocytes
Other
I
CD4+ Tconv cells r=0.60, p=0.002
CD4+ Treg cells r=0.85, p<0.001
CD8+ T cells r=0.61, p=0.002
B cells r=0.83, p<0.001
NK cells r=0.78, p<0.001
CD14+ cells r=0.49, p=0.017
DCs r=0.32, p=0.14
pDCs r=0.24, p=0.27
scRNAseq (% of cells)
Flow cytometry (% of cells)

Figure S2. DRAGON clustering and identification of cell types, related to Figure 1. A model-based clustering algorithm combining deterministic annealing and Gaussian mixture models was developed for clustering of single-cell RNAseq data. (A) DRAGON takes as input an annealing schedule of "temperatures", the principal components embedding of the cells, and the maximum number of clusters. After initialization, the algorithm enters a loop were the temperature is selected from the annealing schedule, clusters are split if the temperature is less than 2 times the eigenvalue of the cluster, and then expectation/maximization based on a Gibbs distribution with temperature as an agitation parameter that seeks to maximize the parameters of the Gaussian mixture model. After convergence or a set number of iterations, the temperature is lowered and the loop begins again to allow the system to anneal. This process prevents the clustering algorithm from becoming trapped in local minima, allows for the assessment of clusters as the temperature decreases, provides a rationale for cluster genesis, and is highly parallelizable for large datasets. (B) Results at decreasing annealing temperatures for all single-cells from our dataset. The strongest separations between clusters are evident at the higher temperatures, and more subtle separations become evident as the system is gradually cooled. (C) To benchmark DRAGON versus Louvian clustering, we generated 50 instances of synthetic data using splatter (STAR Methods). We used a framework where there were 4 clusters, with varying frequencies of cells in each cluster. The top two panels show the frequencies of cells per cluster and the number of differentially expressed genes per cluster for each of the synthetic datasets. We used confusion matrices to compare the accuracy of DRAGON and Louvian clustering versus ground truth (bottom panels), and found that both algorithms were highly accurate (90.6% and 90.3% accuracy, respectively) and correlated (rho=0.90, p<0.001). (D) Differentially expressed genes across clusters from Temperature=45 in (B) revealed separation between lineage-defining markers across immune cells. (E) Sub-clustering of each of the 4 clusters from Temperature=45 in (B) and identification of immune lineages. (F-G) Flow cytometric gating to identify major immune lineages in lymphocytes (F) and myeloid cells (G). (H) Comparison of bioinformatically inferred cell types and populations identified by flow cytometry strongly agreed. (I) Correlations between cell populations identified by flow cytometry versus those inferred bioinformatically.

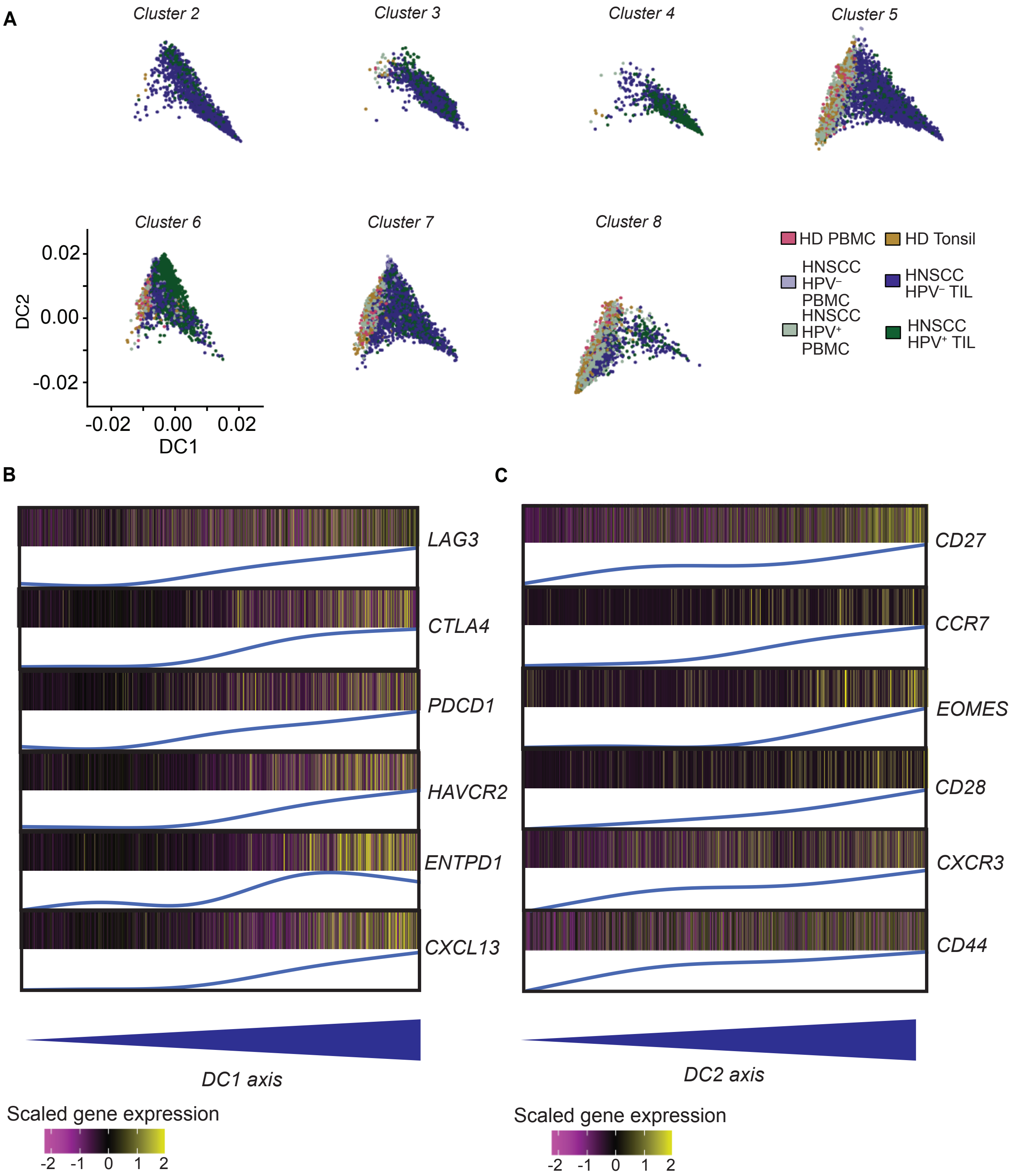

A
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
0.02
0.00
-0.02
DC2
-0.02
0.00
0.02
DC1
HD PBMC
HD Tonsil
HNSCC HPV⁻ PBMC
HNSCC HPV⁻ TIL
HNSCC HPV⁺ PBMC
HNSCC HPV⁺ TIL
B
LAG3
CTLA4
PDCD1
HAVCR2
ENTPD1
CXCL13
DC1 axis
Scaled gene expression
-2 -1 0 1 2
C
CD27
CCR7
EOMES
CD28
CXCR3
CD44
DC2 axis
Scaled gene expression
-2 -1 0 1 2

Figure S3. CD8$^+$ T cells from HPV$^-$ and HPV$^+$ TIL share a differentiation trajectory, related to Figure 2. Diffusion map embedding was performed using clusters 2-8 of CD8$^+$ T cells (cluster 1 was excluded due to high expression of genes associated with the cell cycle). (A) Diffusion map embedding of each cluster of CD8+ T cells demonstrates how each of these clusters relate to the differentiation trajectory. (B) To determine the genes associated with differentiation of CD8$^+$ T cells, we performed linear modeling between gene expression levels and the DC1 axis. Genes for known co-inhibitory receptors such as *LAG3*, *CTLA4*, *PDCD1* (gene for PD1), and *HAVCR2* (gene for TIM3) were highly associated with DC1, as were *ENTPD1* (gene for CD39) and *CXCL13*. (C) We also assessed genes associated with the DC2 axis, and found genes associated with effector memory cells such as *CCR7* (to control trafficking to lymph nodes), *EOMES* (required for memory formation), *CD27* and *CD28* (associated with memory cells).

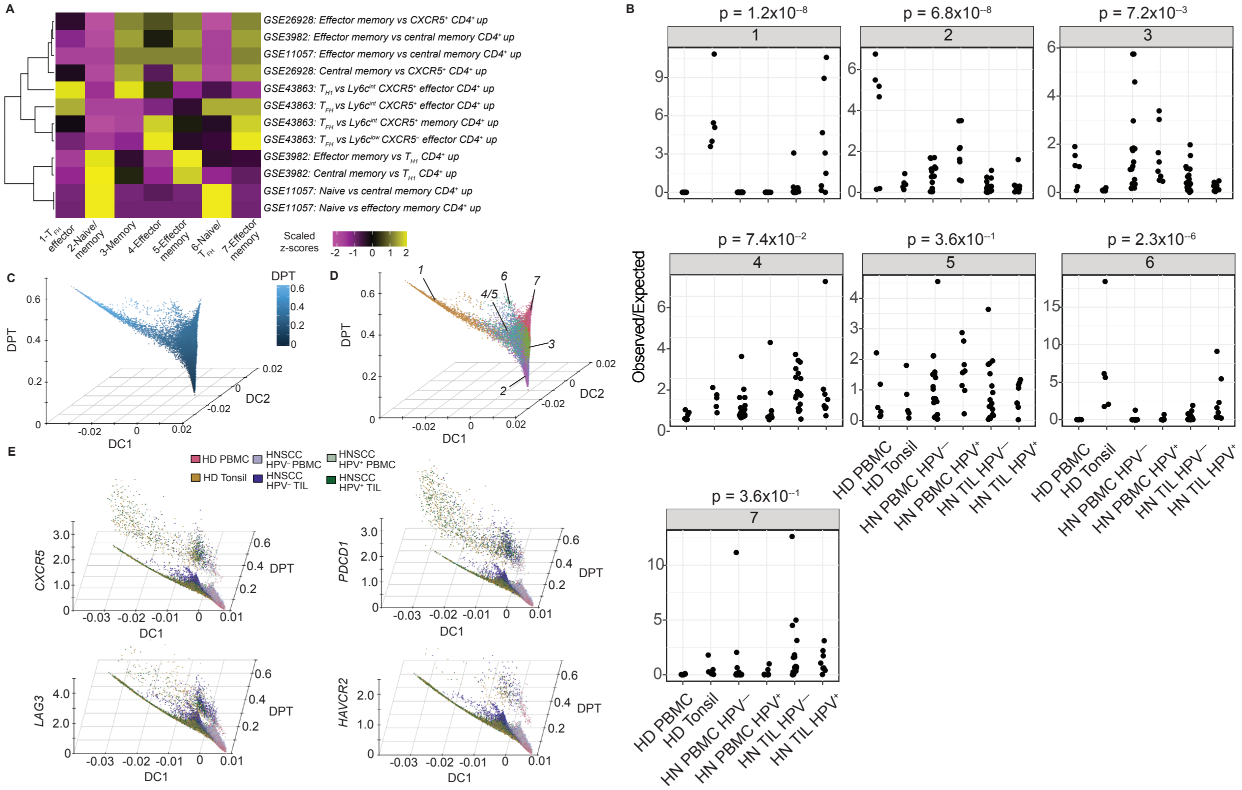


Figure S4. Elucidation of biological roles of CD4$^+$ Tconv clusters, and diffusion map embedding of differentiation trajectories, related to Figure 3. (A) Gene set enrichment analysis of CD4$^+$ Tconv clusters using CD4$^+$ Tconv related gene sets from the C7 collection of MSigDB. Curated gene sets are shown to highlight clusters found to be associated with naïve and memory states (clusters 2 and 6) and various stages of effector, type-1 helper (TH1) and TFH development. (B) Diffusion map embedding with DC1 and DC2 on the x and z axes, and diffusion pseudotime (DPT) on the y axis, colored by diffusion pseudotime (DPT). (C) Same diffusion map embedding as (B), but with colored by clusters from Figure 3A. (D) Diffusion map embeddings of DC1 (x axis), DPT (z axis) and individual gene expression levels (y axis). Expression of *PDCD1* and *CXCR5* is bimodal with respect to DC1 and DPT (top two panels), consistent with expression during early activation and in a differentiated TFH state. Conversely, *LAG3* and *HAVCR2* are co-expressed with *PDCD1* prior to the point were differentiation to TFH occurs, suggesting that co-inhibitory receptor expression may prevent CD4$^+$ Tconv differentiation. (E) To determine whether there were statistically significant associations between clusters and samples types for CD4+ Tconv, we generated a metric for the observed versus expected frequencies of cells in each cluster, where each dot represent a patient from a given sample type (STAR Methods). We then performed ANOVA on each cluster, with the resulting p values displayed above each numbered cluster. The resulted in the identification of statistically significant associations between sample type and clusters 1, 2, 3, and 6. Importantly, we also compared HN TIL HPV^{-} CD4^{+} Tconv and HN TIL HPV^{+} CD4^{+} Tconv for cluster 1, and found that they were statistically significantly different (p=0.0044, rank sum test).

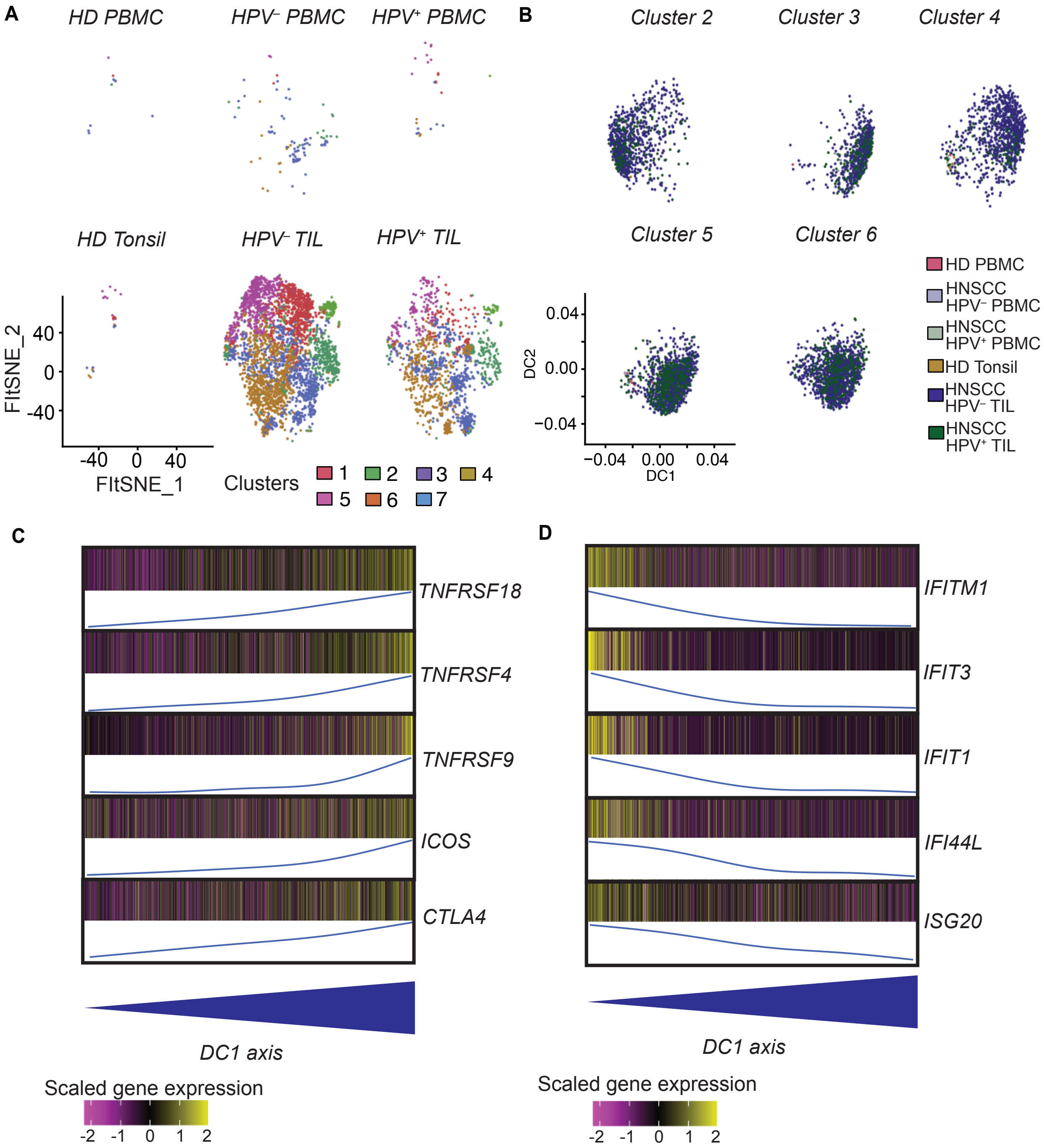

A
HD PBMC
HPV⁻ PBMC
HPV⁺ PBMC
HD Tonsil
HPV⁻ TIL
HPV⁺ TIL
FItSNE_2
40
0
-40
-40
0
40
FItSNE_1
Clusters
1
2
3
4
5
6
7
B
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
DC2
0.04
0.00
−0.04
−0.04
0.00
0.04
DC1
HD PBMC
HNSCC HPV⁻ PBMC
HNSCC HPV⁺ PBMC
HD Tonsil
HNSCC HPV⁻ TIL
HNSCC HPV⁺ TIL
C
TNFRSF18
TNFRSF4
TNFRSF9
ICOS
CTLA4
DC1 axis
Scaled gene expression
-2
-1
0
1
2
D
IFITM1
IFIT3
IFIT1
IFI44L
ISG20
DC1 axis
Scaled gene expression
-2
-1
0
1
2

Figure S5. CD4$^+$ Treg exhibit a conserved differentiation trajectory in HPV$^-$ and HPV$^+$ TIL, related to Figure 3. Pseudotime derived from diffusion map embedding revealed expression of TNF-family receptors following activation. (A) The same FItSNE embedding of CD4$^+$ Treg as in Figure 3, but showing each of the sample types (with the plots colored by clusters, as in Figure 3). (B) Diffusion map showing the locations of each of the clusters identified in Figure 3F. Clusters 2 and 4 are both enriched for IFN-responsive gene sets and are found at early diffusion pseudotime (DPT), while clusters 3 and 6 are enriched for TNF-family receptor genes and are found at later DPT. Cluster 5 is potentially a transient state linking the IFN-responsive and TNFR-superfamily states. (C) Genes positively associated with the DC1 component of the diffusion map embedding. Notably, TNF-family receptors, co-stimulatory molecules and Treg effector molecules are associated with DC1. (D) Conversely, top genes that are negatively associated with DC1 include members of the IFN-response gene family.

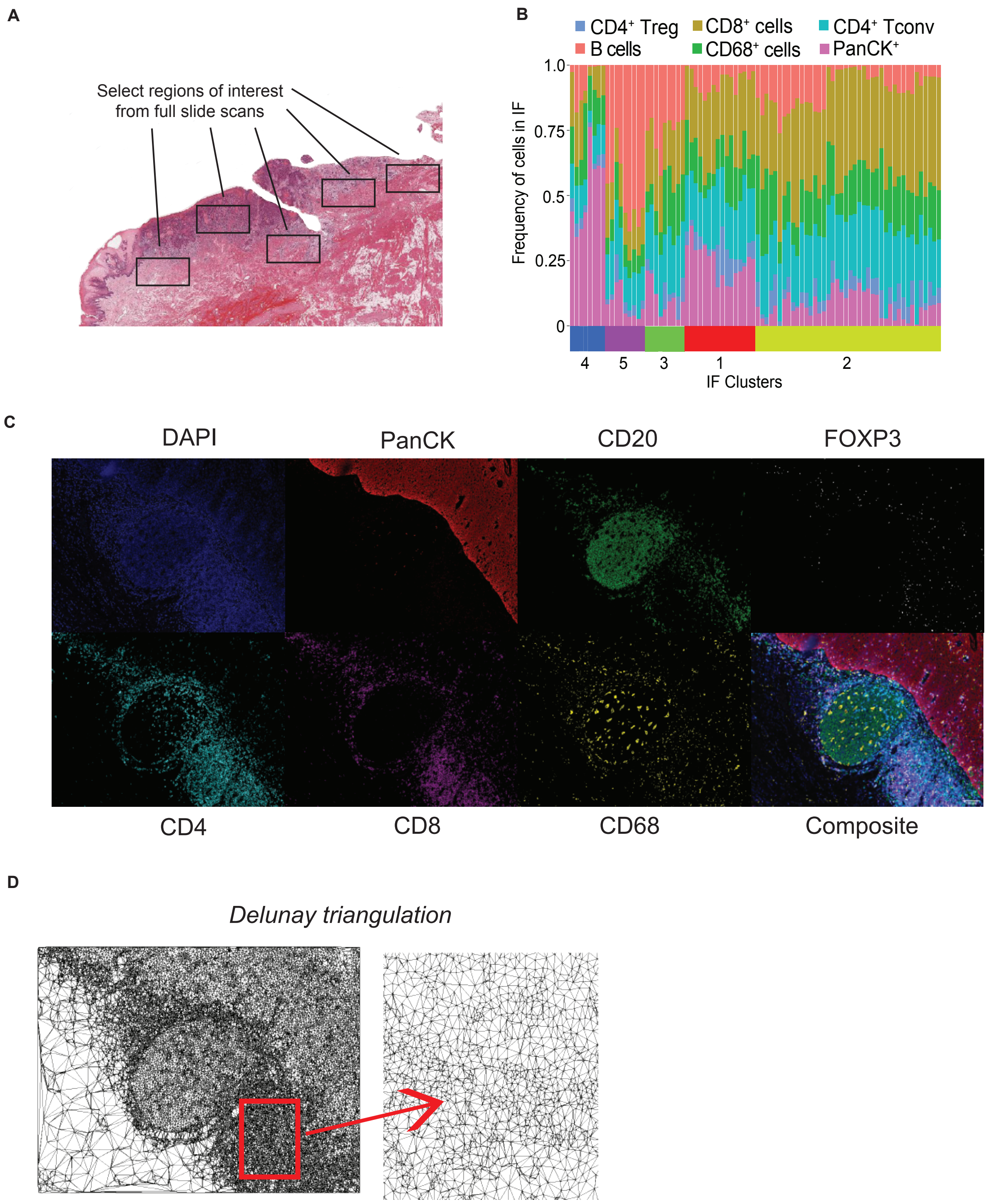
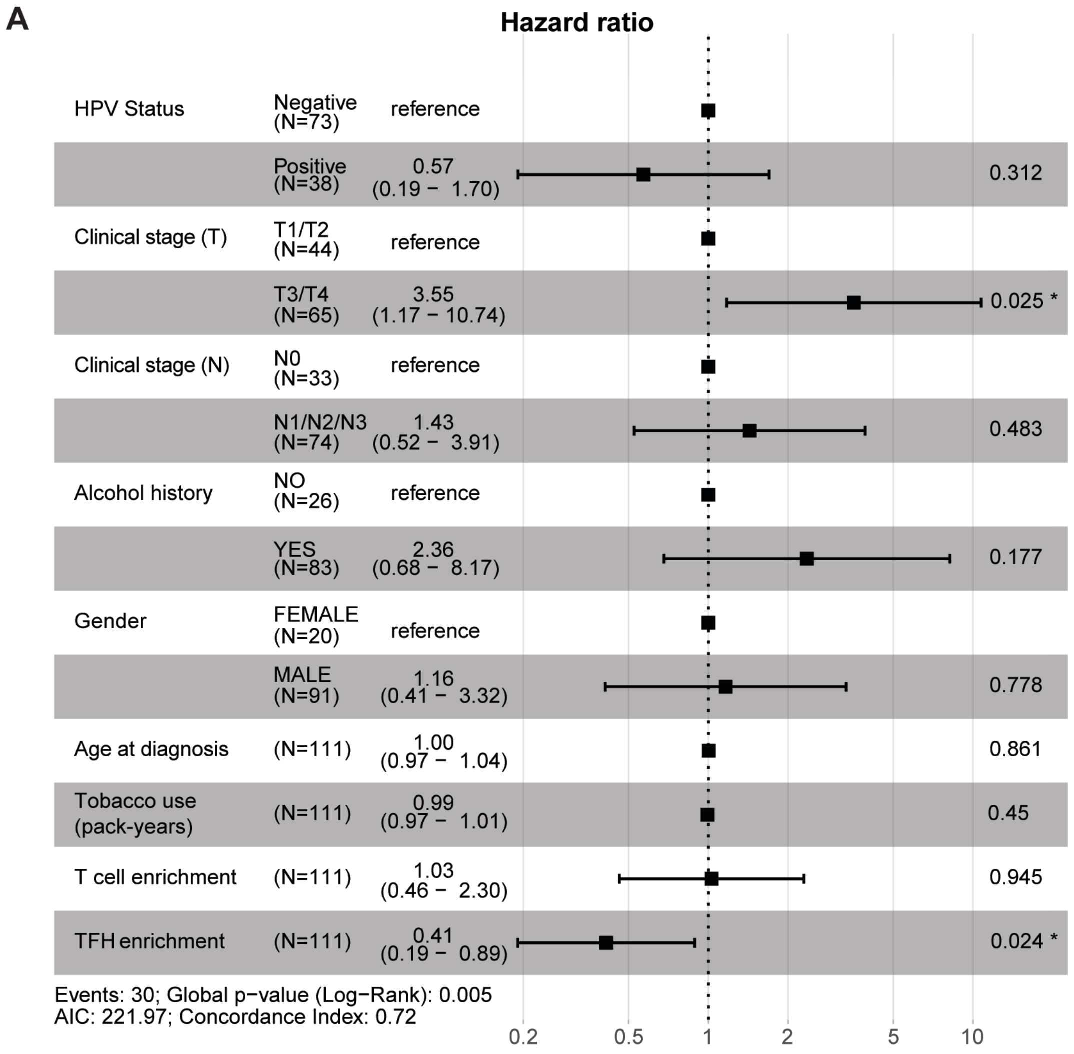

A
Select regions of interest from full slide scans
B
🟦 CD4⁺ Treg
🟨 CD8⁺ cells
🟦 CD4⁺ Tconv
🟥 B cells
🟩 CD68⁺ cells
🟪 PanCK⁺
Frequency of cells in IF
1.0
0.75
0.5
0.25
0
4
5
3
1
2
IF Clusters
C
DAPI
PanCK
CD20
FOXP3
CD4
CD8
CD68
Composite
D
Delunay triangulation

Figure S6. Representative immunofluorescence data, identification of cell type frequencies, and workflow for log odds ratio of cell-cell neighbors, related to Figure 7. Immunofluoresence analysis was used to characterize spatial relationships in tissue samples. (A) Regions of interest were selected based on H&E staining in addition to whole slide scans. (B) Quantification of cell types as a frequency of potential cell types identifiable by our IF panel. The clustering as a result of the frequencies of cells in a given image (derived from Figure 7A) is shown on the x axis. (C) Representative individual channels and composite images from the immunofluorescence workflow. (D) Representative Delunay triangulation analysis from the slide in (C) on the entire image (left) and a zoomed in section of the image (right). Delunay triangulation was used to determine the neighbors of a given cell type in 2 dimensional images.



HPV Status	Negative (N=73)	reference	
	Positive (N=38)	0.57 (0.19 − 1.70)	0.312
Clinical stage (T)	T1/T2 (N=44)	reference	
	T3/T4 (N=65)	3.55 (1.17 − 10.74)	0.025 *
Clinical stage (N)	N0 (N=33)	reference	
	N1/N2/N3 (N=74)	1.43 (0.52 − 3.91)	0.483
Alcohol history	NO (N=26)	reference	
	YES (N=83)	2.36 (0.68 − 8.17)	0.177
Gender	FEMALE (N=20)	reference	
	MALE (N=91)	1.16 (0.41 − 3.32)	0.778
Age at diagnosis	(N=111)	1.00 (0.97 − 1.04)	0.861
Tobacco use (pack-years)	(N=111)	0.99 (0.97 − 1.01)	0.45
T cell enrichment	(N=111)	1.03 (0.46 − 2.30)	0.945
TFH enrichment	(N=111)	0.41 (0.19 − 0.89)	0.024 *

Events: 30; Global p−value (Log−Rank): 0.005
AIC: 221.97; Concordance Index: 0.72

Figure S7. Multivariate progression free survival analysis shows TFH enrichment remains statistically significant after correcting for potentially confounding covariates, related to Figure 7. (A) Multivariate analysis of progression free survival by covariates and TFH enrichment. Notably, TFH enrichment remained a statistically significant independent predictor of longer progression free interval in this multivariate analysis.