Single cell antigen receptor analysis reveals lymphocyte developmental origins

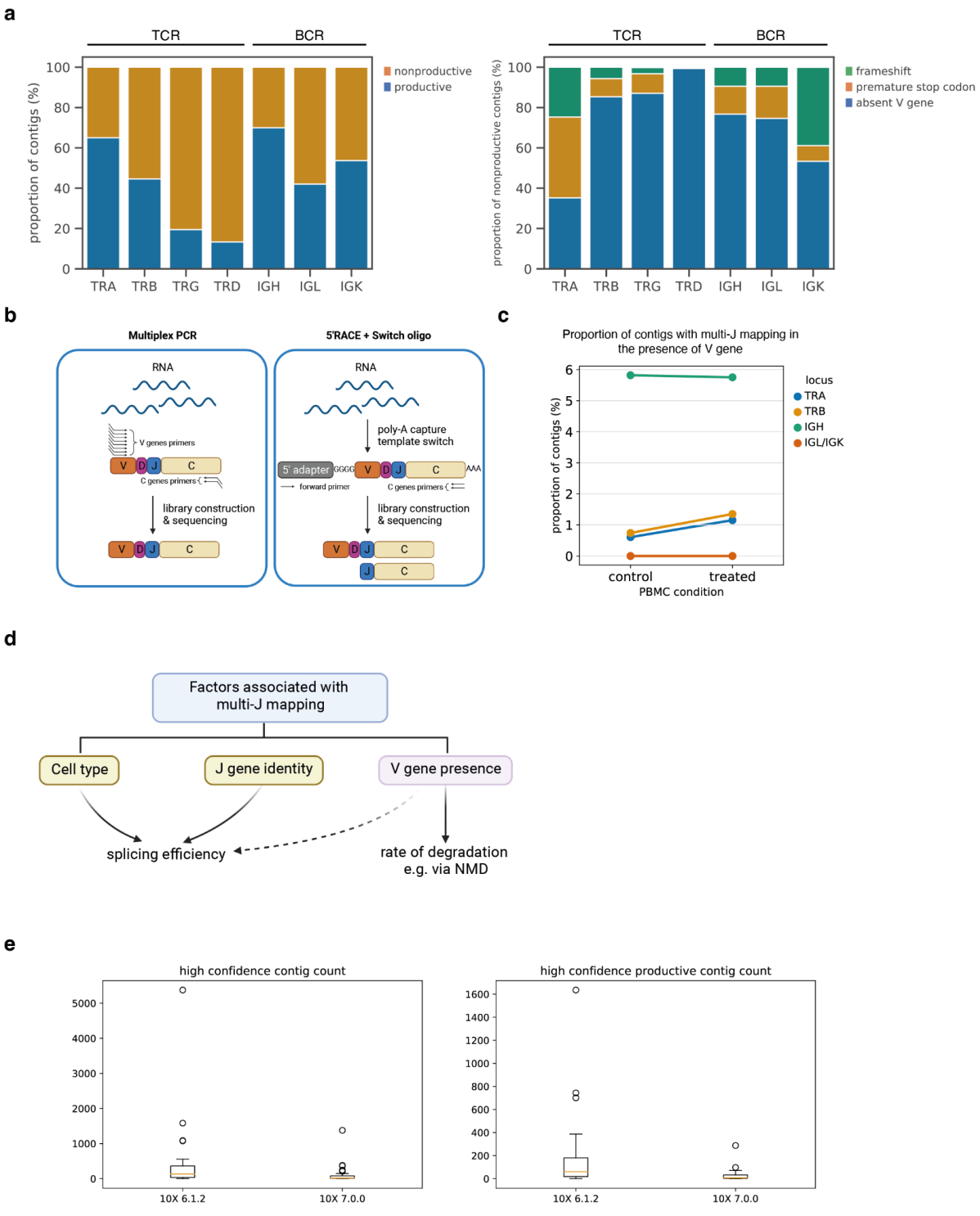
Authors: Chenqu Suo[1,2,8], Krzysztof Polanski[1,8], Emma Dann[1], Rik G.H. Lindeboom[1], Roser Vilarrasa Blasi[1], Roser Vento-Tormo[1], Muzlifah Haniffa[1,3,4], Kerstin B. Meyer[1], Zewen Kelvin Tuong[1,5,7,9]*, Menna R. Clatworthy[1,5,9]*, Sarah A. Teichmann[1,6,9]*

Affiliations:
[1]Wellcome Sanger Institute; Wellcome Genome Campus, Hinxton, Cambridge, UK.
[2]Department of Paediatrics, Cambridge University Hospitals; Hills Road, Cambridge, UK.
[3]Biosciences Institute, Newcastle University; Newcastle upon Tyne, UK.
[4]Department of Dermatology and NIHR Newcastle Biomedical Research Centre, Newcastle upon Tyne Hospitals NHS Foundation Trust; Newcastle upon Tyne, UK.
[5]Molecular Immunity Unit, University of Cambridge Department of Medicine; Cambridge, UK.
[6]Theory of Condensed Matter, Cavendish Laboratory, Department of Physics, University of Cambridge; Cambridge, UK.
[7]Frazer Institute, Faculty of Medicine, The University of Queensland, Brisbane, Australia
[8]These authors contributed equally to this work.
[9]These senior authors contributed equally to this work.
*Corresponding authors. Email: zkt22@cam.ac.uk (Z.K.T.), mrc38@cam.ac.uk (M.R.C.), st9@sanger.ac.uk (S.A.T.).

Supplementary Figures

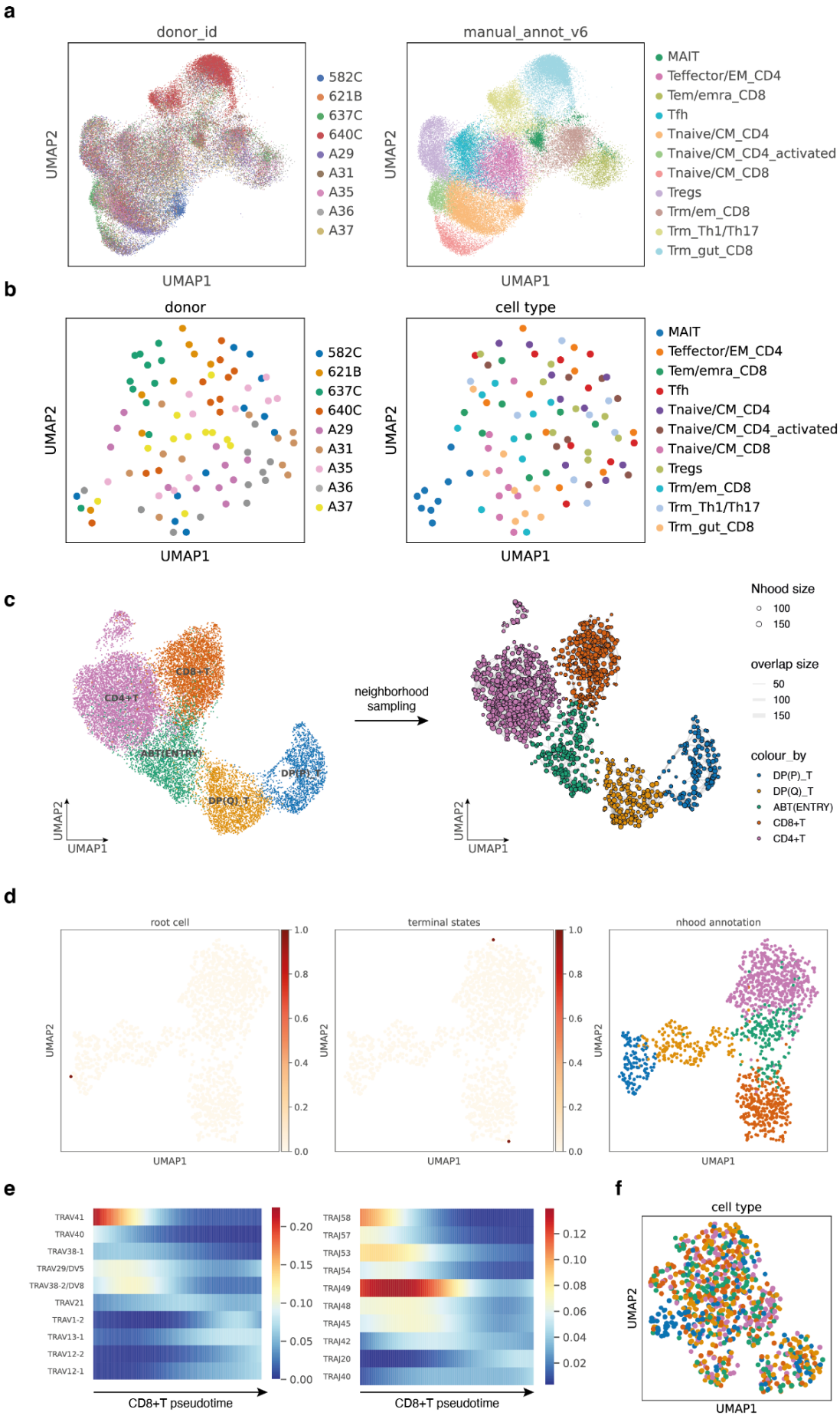
	Immcantation Framework	enclone	scRepertoire	VDJView	Immunarch	scirpy	dandelion
Programming Language	R & Python	Rust	R	R	R	Python	Python
Version Control	Bitbucket	Github	Github	Bitbucket	Github	Github	Github
AIRR Sofware Certified	✓	✗	✗	✗	✗	✓	✓
TCR/BCR Centric	BCR	BCR	Both	Both	Both	Both	Both
V(D)J Re-annotation	✓	✗	✗	✗	✗	✗	✓
Clone Definition	✓	✓	✓	✗	✗	✓	✓
BCR Mutation Quantification	✓	✗	✗	✗	✗	✗	Through Immcantation
Diversity Estimation	✓	✗	✓	✗	✓	✓	✓
Visualization	Minimal	✓	✓	✓	✓	✓	✓
Single-cell Integration	Minimal	✓	✓	✓	✗	✓	✓
Phylogenetic Lineage Inference	✓	✓	✗	✗	✗	✗	✗
Trajectory Inference	✗	✗	✗	Through Monocle2 (GEX only)	✗	✗	✓

Extended Data Fig. 1 | List of features included in AgR repertoire analysis pipelines. A table outlining the features of other methods compared to *Dandelion*. As the output from *Dandelion* is compatible with any AIRR-compliant softwares e.g. *Dandelion* output can be passed to *Immcantation* to perform phylogenetic lineage inference.

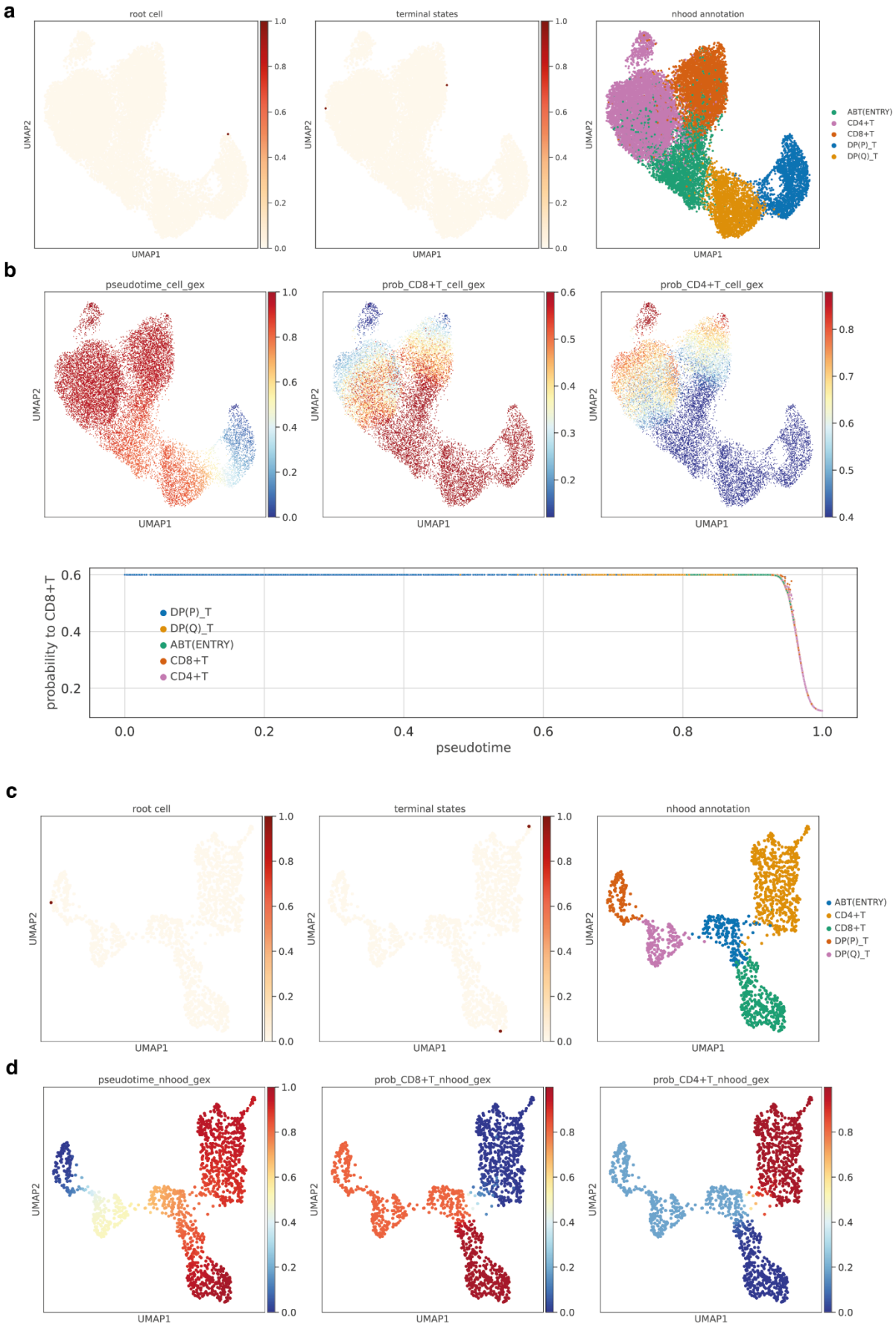


Extended Data Fig. 2 | *Dandelion* offers improved contig annotations. a, Left: barplot of proportion of contigs that are productive or non-productive in each locus. Right: barplot showing the causes of non-productive contigs in each locus. For both plots, sc-γδTCR, -αβTCR and -BCR data were taken from Suo et al. 2022[3] excluding thymus samples. **b,** Schematic illustration showing that mRNA without V genes would be captured by 5′RACE + Switch oligo technique but not by multiplex PCR strategy. **c,** Pointplot of proportion of contigs with multi-J mapping in the presence of V gene in control and cycloheximide-treated PBMC samples. Points are colored by locus of TCR/BCR. **d,** Schematic illustration showing

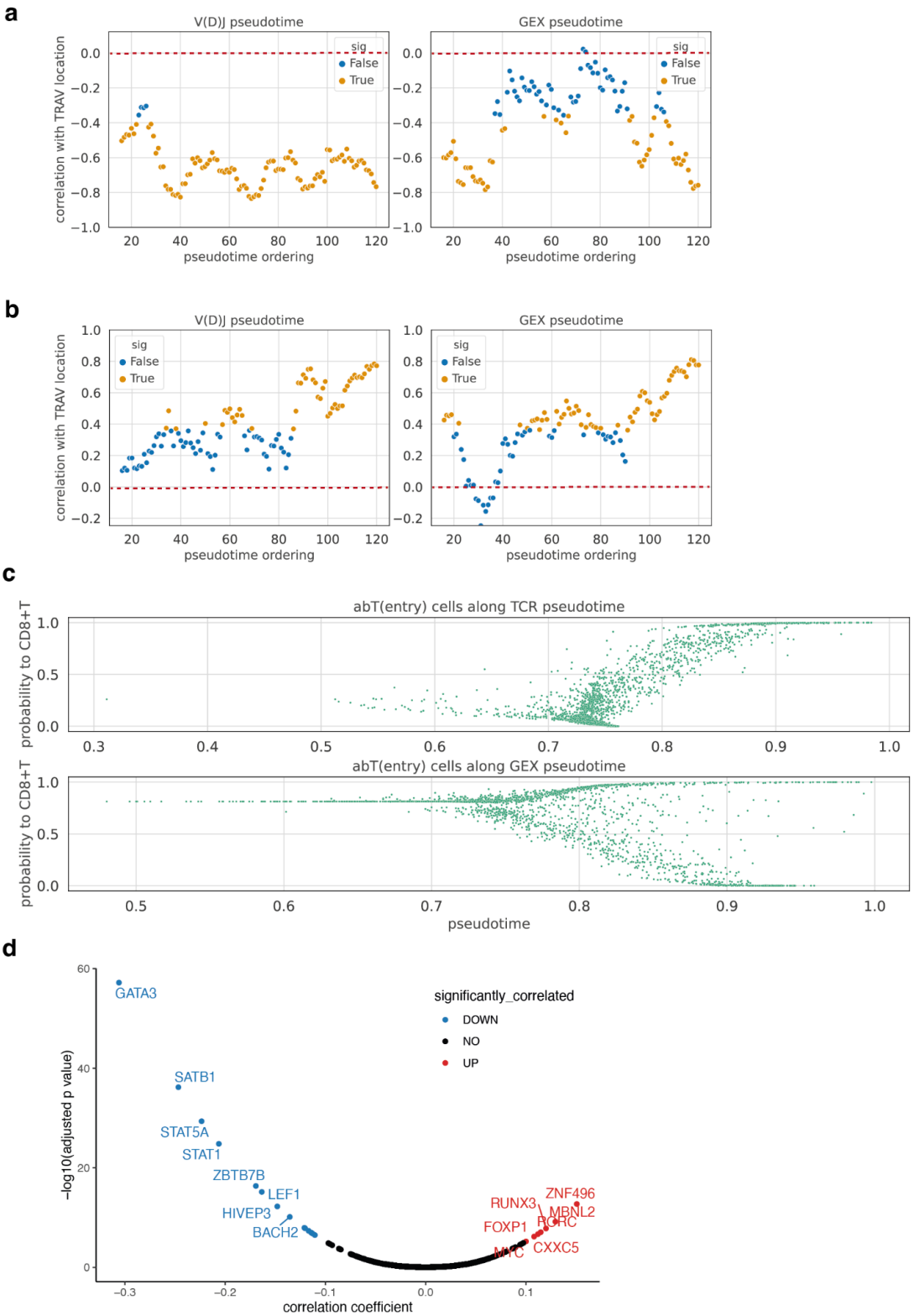
the factors associated with multi-J mapping and the proposed mechanisms. **e,** Boxplots of sc-γδTCR contig counts annotated by 10X *cellranger vdj* v6.1.2 *versus* v7.0.0 using data from Suo et al. 2022[3]. Left: all high confidence contigs (*P*-value 5.43e-6, r 0.91 in the Wilcoxon signed-rank test). Right: high confidence productive contigs (*P*-value 1.69e-6, r 0.96 in the Wilcoxon signed-rank test).


a
donor_id
582C
621B
637C
640C
A29
A31
A35
A36
A37
UMAP1
UMAP2
manual_annot_v6
MAIT
Teffector/EM_CD4
Tem/emra_CD8
Tfh
Tnaive/CM_CD4
Tnaive/CM_CD4_activated
Tnaive/CM_CD8
Tregs
Trm/em_CD8
Trm_Th1/Th17
Trm_gut_CD8
b
donor
cell type
c
CD8+T
CD4+T
ABT(ENTRY)
DP(P)_T
DP(Q)_T
neighborhood sampling
Nhood size
100
150
overlap size
50
100
150
colour_by
d
root cell
terminal states
nhood annotation
e
TRAV41
TRAV40
TRAV38-1
TRAV29/DV5
TRAV38-2/DV8
TRAV21
TRAV1-2
TRAV13-1
TRAV12-2
TRAV12-1
CD8+T pseudotime
TRAJ58
TRAJ57
TRAJ53
TRAJ54
TRAJ49
TRAJ48
TRAJ45
TRAJ42
TRAJ20
TRAJ40
f

Extended Data Fig. 3 | V(D)J feature space. a, Gene expression UMAP of all T cells from Conde et al. 2022[5], colored by donor ID (left) or high-level cell type annotations (right). Each point represents a cell. **b,** UMAP of the pseudo-bulk V(D)J feature space of the same cells as in **a**, colored by donor ID (left) or high-level cell type annotations (right). Each point represents a cell pseudo-bulk. **c,** Left: UMAP of DP to mature T cells with paired productive αβTCR in data from Suo et al. 2022[3]. Each point represents a cell, colored by cell types. Right: cell neighborhood graph on the same UMAP embedding. Each point represents a cell neighborhood, colored by cell types. The point size represents neighborhood size, with connecting edges representing overlapping cell numbers between any two neighborhoods. Only edges with more than 30 overlapping cells are shown. The layout of nodes is determined by the position of the neighborhood index cell in the UMAP on the left. **d,** The root cell and terminal states selected for pseudotime inference in **Fig. 3c. e,** Gene expression trends over CD8+T pseudotime imputed with *palantir*[23]. Only the top 10 most frequently used TRAV or TRAJ genes are shown. **f,** UMAP representation of tcrdist-derived PCA coordinates of VDJ data computed by CoNGA[24], with the same dataset as used in **c**, colored by cell types.

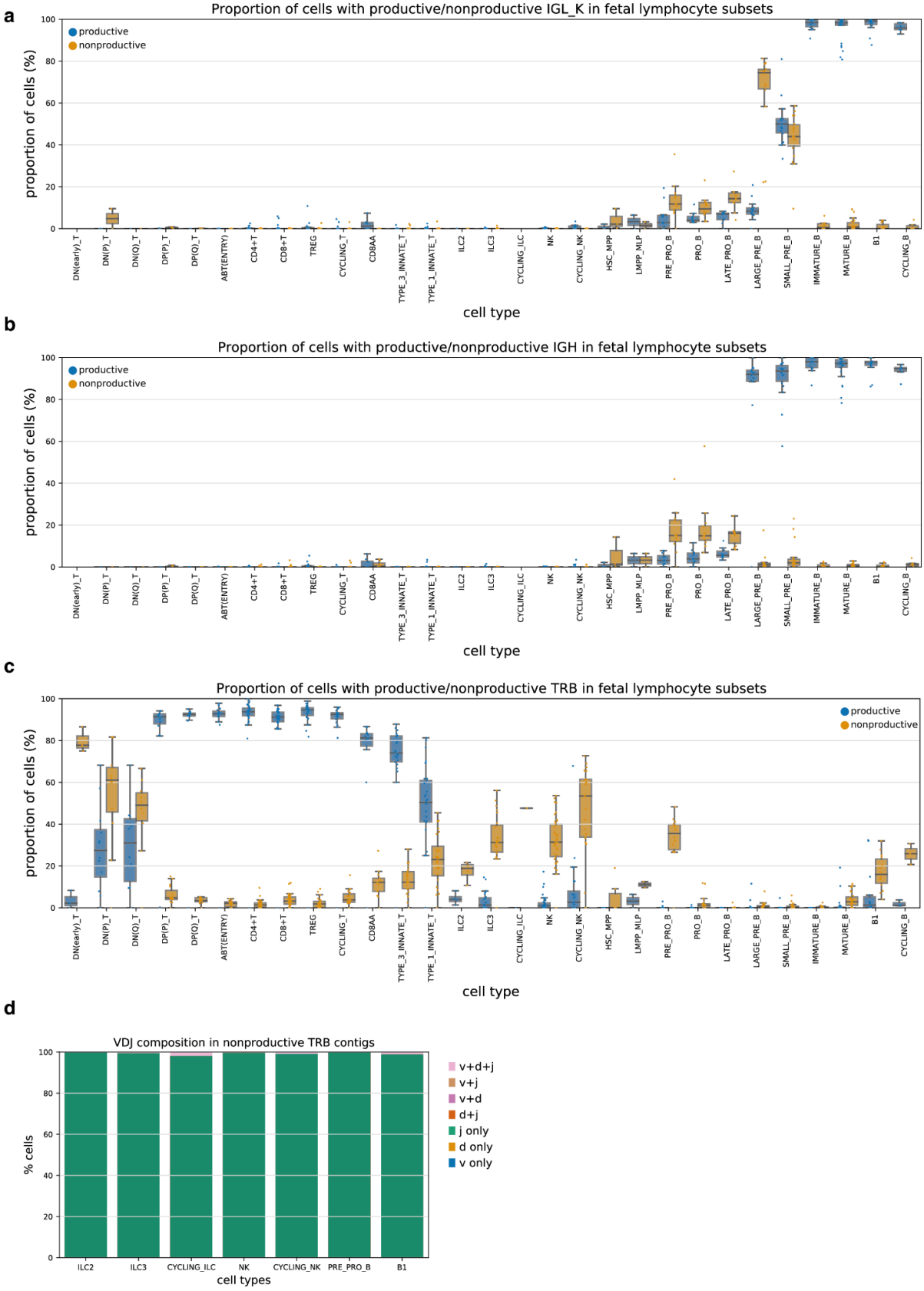

a
root cell
terminal states
nhood annotation
UMAP1
UMAP2
ABT(ENTRY)
CD4+T
CD8+T
DP(P)_T
DP(Q)_T
b
pseudotime_cell_gex
prob_CD8+T_cell_gex
prob_CD4+T_cell_gex
probability to CD8+T
pseudotime
c
root cell
terminal states
nhood annotation
d
pseudotime_nhood_gex
prob_CD8+T_nhood_gex
prob_CD4+T_nhood_gex

Extended Data Fig. 4 | T cell development pseudotime inference comparison. a, DP to mature T cells with paired productive αβTCR in data from Suo et al. 2022[3], on the same UMAP embedding as in **Fig. 4a** and **Supplementary Fig. 3c**. The first two panels show the root cell and terminal states selected for pseudotime inferred directly from single-cell gene expression. The last panel shows the cell types. **b,** Top: pseudotime and branch probabilities inferred directly from single-cell gene expression on the same UMAP embedding as in **a**. Bottom: scatterplot of branch probability to CD8+T against pseudotime. Each point represents a cell. **c,** UMAP of neighborhood GEX space, with the same neighborhoods as sampled in **Supplementary Fig. 3c** and UMAP embedding computed on gene expression pseudo-bulked by neighborhoods. Each point represents a cell neighborhood. The first two panels show the root cell and terminal states selected for pseudotime inferred from neighborhood GEX space. The last panel shows the cell types. **d,** Inferred pseudotime, and branch probabilities to CD8+T and to CD4+T respectively overlaid onto the same UMAP embedding in **c**.



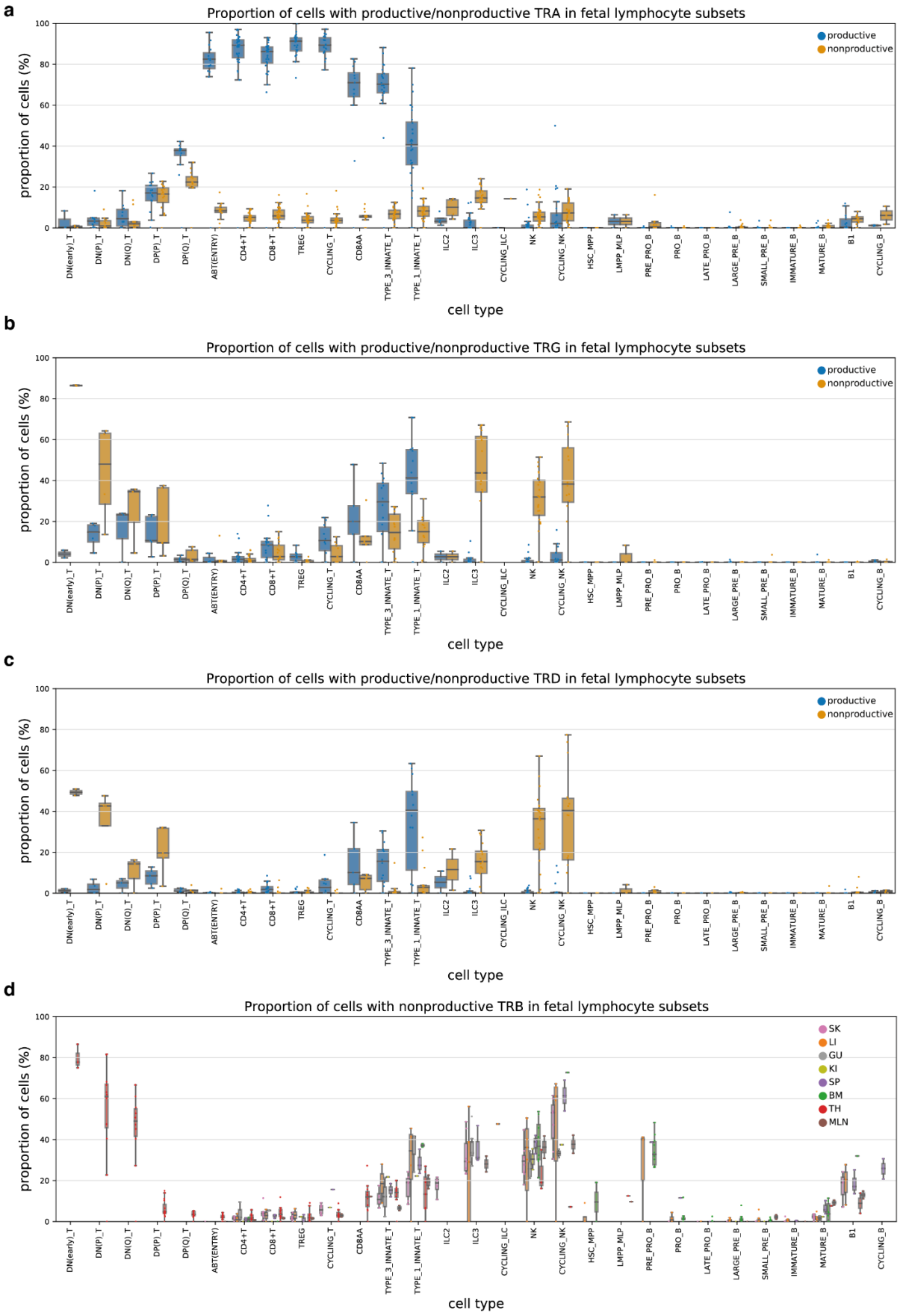
Extended Data Fig. 5 | Comparing pseudotime inferred from neighborhood V(D)J space or GEX space. a, Pearson's correlation coefficients of pseudotime order and average relative TRAV location over sliding windows of 30 adjacent neighborhoods on the pseudotime order (left: pseudotime inferred from neighborhood V(D)J space; right: pseudotime inferred from neighborhood GEX space). *Y*-axis is the correlation coefficient and the x-axis is the median pseudotime order of the 30 adjacent neighborhoods. The color of the points represents

statistical significance (orange: *P*-value from the Pearson's correlation < 0.05; blue: *P*-value ≥ 0.05). The red dashed lines mark the correlation coefficient of 0. **b,** The same plot as in **a** but for TRAJ. **c,** Scatterplots of branch probability to CD8+T against pseudotime in abT(entry) cells. Each point represents a cell. Top panel: pseudotime inferred from neighborhood V(D)J space as in **Fig. 4a** top panel. Bottom panel: pseudotime inferred from neighborhood GEX space as in **Fig. 4a** bottom right panel. **d,** Volcano plot summarizing results of TFs that are correlated with branch probabilities to CD8+T lineage in V(D)J pseudotime within abT(entry) cells. The *y*-axis is the $-\log_{10}$(BH adjusted *P*-value) and the *x*-axis is the correlation coefficient. Labeled TFs that had significant (BH adjusted *P*-value < 0.05) positive correlations (correlation coefficient > 0.1) were colored in red, the ones with significant negative correlations (correlation coefficient < -0.1) were colored in blue, and the rest were colored in black.

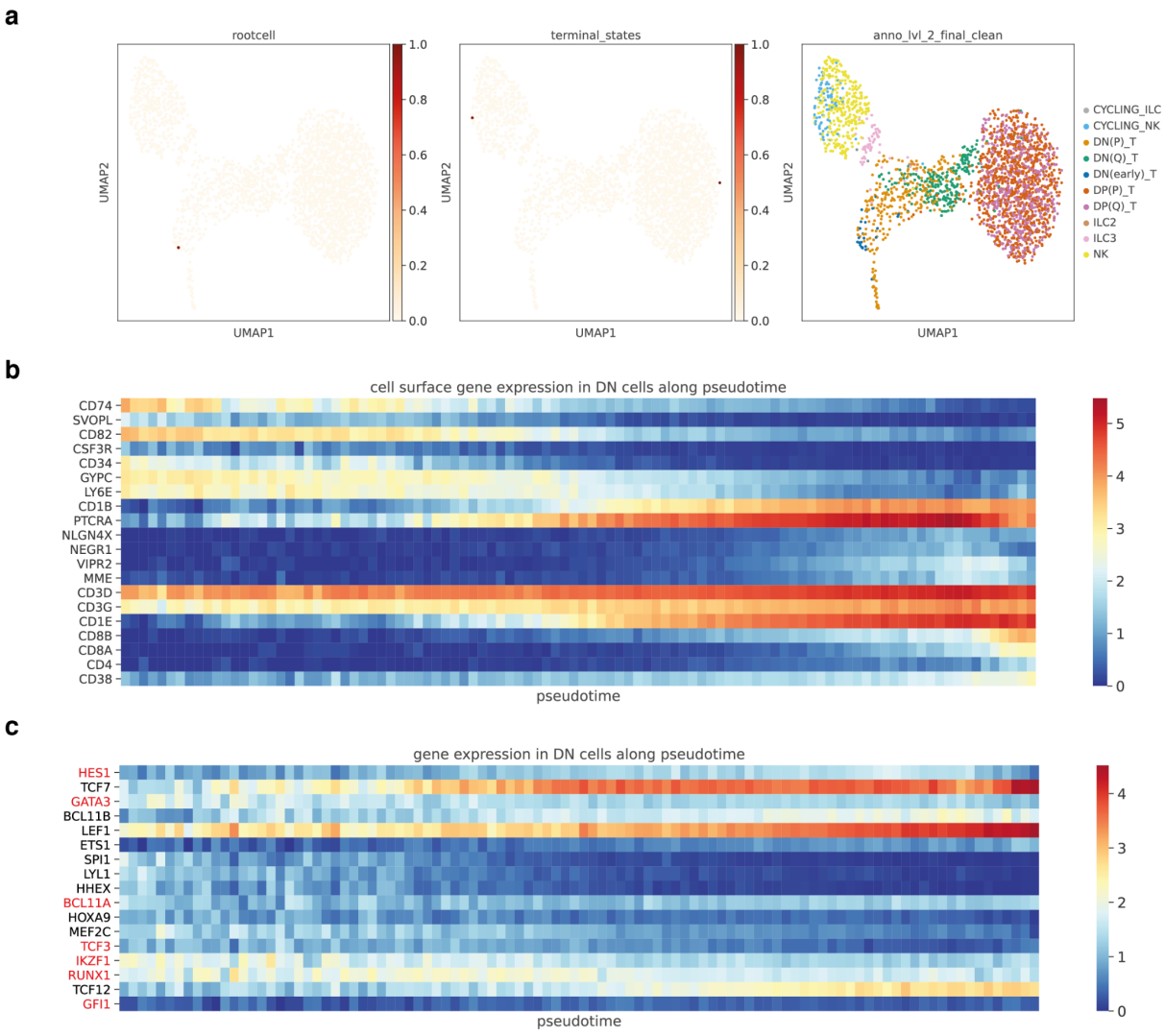

a
Proportion of cells with productive/nonproductive IGL_K in fetal lymphocyte subsets
productive
nonproductive
proportion of cells (%)
cell type
b
Proportion of cells with productive/nonproductive IGH in fetal lymphocyte subsets
productive
nonproductive
proportion of cells (%)
cell type
c
Proportion of cells with productive/nonproductive TRB in fetal lymphocyte subsets
productive
nonproductive
proportion of cells (%)
cell type
d
VDJ composition in nonproductive TRB contigs
v+d+j
v+j
v+d
d+j
j only
d only
v only
% cells
ILC2
ILC3
CYCLING_ILC
NK
CYCLING_NK
PRE_PRO_B
B1
cell types

1

Extended Data Fig. 6 | Non-productive BCR and TCR. a,b,c, Boxplot of the proportion of cells with productive (blue) or non-productive (orange) BCR light chain (**a**) and heavy chain (**b**), and TRB (**c**) in different fetal lymphocyte subsets. Each point represents a sample and data were taken from Suo et al. 2022[3]. Only samples with at least 20 cells are shown. Boxes capture the first to third quartiles and whisks span a further 1.5X interquartile range on each side of the box. **d,** Barplot showing the VDJ composition of non-productive TRB contigs in selected lymphocyte subsets from **Fig. 5a**.


a
Proportion of cells with productive/nonproductive TRA in fetal lymphocyte subsets
b
Proportion of cells with productive/nonproductive TRG in fetal lymphocyte subsets
c
Proportion of cells with productive/nonproductive TRD in fetal lymphocyte subsets
d
Proportion of cells with nonproductive TRB in fetal lymphocyte subsets
proportion of cells (%)
cell type
productive
nonproductive
SK
LI
GU
KI
SP
BM
TH
MLN
DN(early)_T
DN(P)_T
DN(Q)_T
DP(P)_T
DP(Q)_T
ABT(ENTRY)
CD4+T
CD8+T
TREG
CYCLING_T
CD8AA
TYPE_3_INNATE_T
TYPE_1_INNATE_T
ILC2
ILC3
CYCLING_ILC
NK
CYCLING_NK
HSC_MPP
LMPP_MLP
PRE_PRO_B
PRO_B
LATE_PRO_B
LARGE_PRE_B
SMALL_PRE_B
IMMATURE_B
MATURE_B
B1
CYCLING_B

Extended Data Fig. 7 | Non-productive TCR. a,b,c, Boxplot of the proportion of cells with productive (blue) or non-productive (orange) TRA (**a**), TRG (**b**) and TRD (**c**) in different fetal lymphocyte subsets. Each point represents a sample and data were taken from Suo et al. 2022[3]. Only samples with at least 20 cells are shown. Boxes capture the first to third quartiles and whisks span a further 1.5X interquartile range on each side of the box. **d,** Boxplot of the proportion of cells with non-productive TRB in different fetal lymphocyte subsets, colored by organs. Each point represents a sample. Only samples with at least 20 cells are shown. Boxes capture the first to third quartiles and whisks span a further 1.5X interquartile range on each side of the box.



Extended Data Fig. 8 | TRBJ-based trajectory for ILC/NK/T cell lineage. a, Neighborhood V(D)J feature space covering ILC, NK and developing T cells with TRBJ on the same UMAP embedding as in **Fig. 5b**. The first two panels show the root cell and terminal states selected for pseudotime inference. The last panel shows the cell types. **b,** Heatmap of gene expression for genes encoding cell surface proteins across pseudotime in DN T cells. Pseudotime is equally divided into 100 bins, and the average gene expression is calculated for DN T cells with pseudotime that falls within each bin. Genes selected here had significantly high Chatterjee's correlation with pseudotime (BH adjusted *P*-value < 0.05, and correlation coefficient > 0.1). **c,** Heatmap of gene expression for TFs known to be important in mouse DN T cell development[40], across pseudotime in human fetal DN T cells. TFs that showed discordant expression patterns between mouse and human are highlighted in red.

Supplementary Tables

Supplementary Table 1: top_10_j_multimappers.csv (separate file)
Top 10 J gene combinations with multi-J mapping for each locus in data from Suo et al. 2022[3], with the number of contigs containing each combination shown next to it.

Supplementary Table 2: LR_results.csv (separate file)
Logistic regression results exploring factors associated with multi-J mapping presence in data from Suo et al. 2022[3].

Supplementary Table 3: LR_results_combined.csv (separate file)
Logistic regression results exploring factors associated with multi-J mapping presence in control and cycloheximide-treated PBMC data.

Supplementary Table 4: j_sequence_affect_j_multimapper.csv (separate file)
List of leftmost (5′ end) J genes that had significant association with increased or decreased multi-J mapping, together with the sequences of their last 10 nucleotides at 3′ ends and the first 11 nucleotides of its 3′ end intron.

Supplementary Table 5: panimmune_differential_VDJ.csv (separate file)
Differential V(D)J usage across CD4+T, CD8+T, and MAIT cells in data from Conde et al. 2022[5].

Supplementary Table 6: abtentry_cor_result.csv (separate file)
Pearson's correlation coefficients and BH adjusted *P*-values of all genes with branch probabilities to CD8+T lineage within abT(entry) cells.
[cor_tcr] Pearson's correlation coefficients for pseudotime inferred from neighborhood V(D)J space
[pval_tcr] Pearson's correlation *P*-values for pseudotime inferred from neighborhood V(D)J space
[adjp_tcr] *P*-values from pval_tcr adjusted by BH procedure
[cor_gex] Pearson's correlation coefficients for pseudotime inferred from neighborhood GEX space
[pval_gex] Pearson's correlation *P*-values for pseudotime inferred from neighborhood GEX space
[adjp_gex] *P*-values from pval_gex adjusted by BH procedure