

UC Office of the President

Research Grants Program Office (RGPO) Funded Publications

Title

Exploration of Cell Development Pathways through High-Dimensional Single Cell Analysis in Trajectory Space

Permalink

<https://escholarship.org/uc/item/3682x5d2>

Journal

iScience, 23(2)

ISSN

2589-0042

Authors

Dermadi, Denis
Bscheider, Michael
Bjegovic, Kristina
et al.

Publication Date

2020-02-01

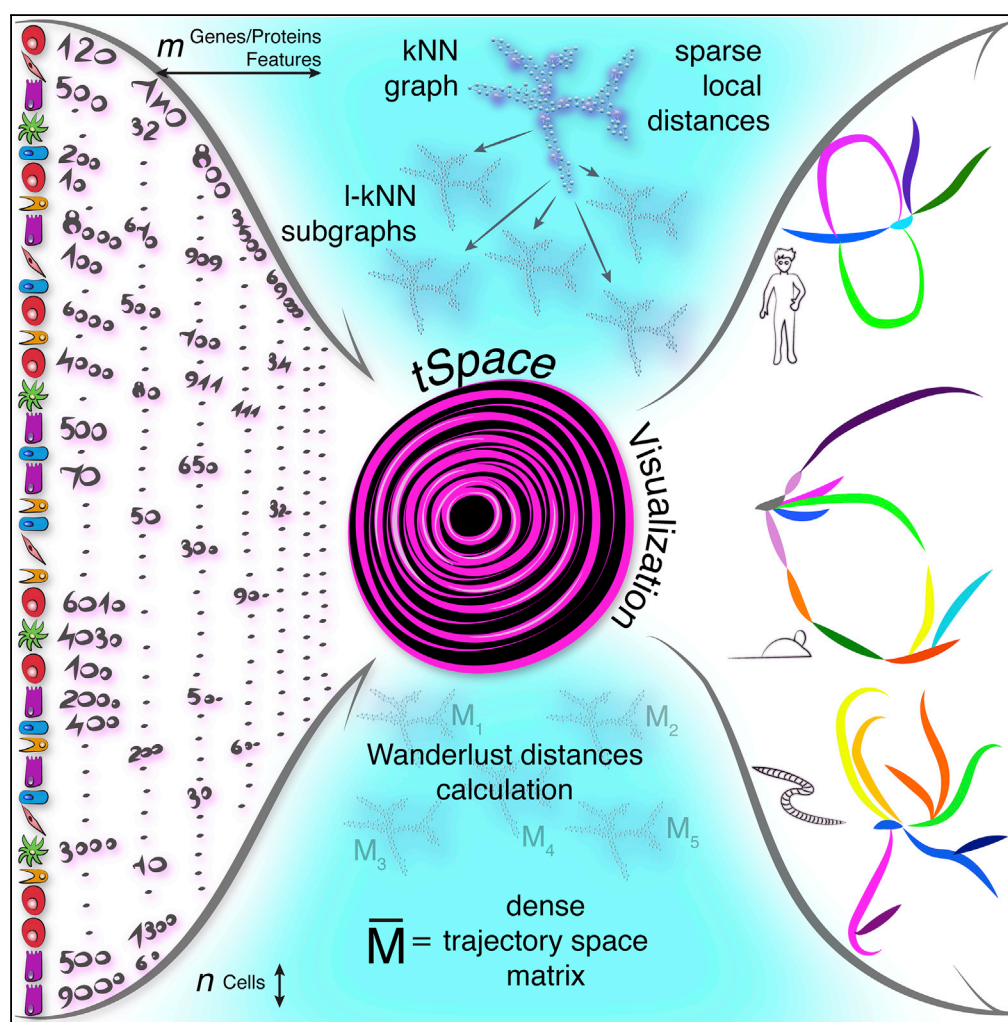
DOI

10.1016/j.isci.2020.100842

Peer reviewed

Article

Exploration of Cell Development Pathways through High-Dimensional Single Cell Analysis in Trajectory Space



Denis Dermadi,
Michael
Bscheider, Kristina
Bjegovic, Nicole
H. Lazarus, Agata
Szade, Husein
Hadeiba, Eugene
C. Butcher

ddermadi@stanford.edu (D.D.)
ebutcher@stanford.edu
(E.C.B.)

HIGHLIGHTS

tSpace: trajectory analysis
agnostic to biological
systems and technology
platforms

tSpace faithfully
reconstructs complex
developmental trees
within one or more tissues

tSpace reveals rare
transient cells usually
missed by trajectory
inference algorithms

tSpace confirms known
and unveils novel biology
in human B cells and
mouse intestine

DATA AND CODE

AVAILABILITY
GSE126954

Dermadi et al., iScience 23,
100842
February 21, 2020 © 2020 The
Author(s).
[https://doi.org/10.1016/
j.isci.2020.100842](https://doi.org/10.1016/j.isci.2020.100842)

Article

Exploration of Cell Development Pathways through High-Dimensional Single Cell Analysis in Trajectory Space

Denis Dermadi,^{1,2,3,4,*} Michael Bscheider,^{1,2,3} Kristina Bjegovic,² Nicole H. Lazarus,^{1,2} Agata Szade,^{1,2} Husein Hadeiba,² and Eugene C. Butcher^{1,2,*}

SUMMARY

High-dimensional single cell profiling coupled with computational modeling is emerging as a powerful tool to elucidate developmental programs directing cell lineages. We introduce tSpace, an algorithm based on the concept of “trajectory space”, in which cells are defined by their distance along nearest neighbor pathways to every other cell in a population. Graphical mapping of cells in trajectory space allows unsupervised reconstruction and exploration of complex developmental sequences. Applied to flow and mass cytometry data, the method faithfully reconstructs thymic T cell development and reveals development and trafficking regulation of tonsillar B cells. Applied to the single cell transcriptome of mouse intestine and *C. elegans*, the method recapitulates development from intestinal stem cells to specialized epithelial phenotypes more faithfully than existing algorithms and orders *C. elegans* cells concordantly to the associated embryonic time. tSpace profiling of complex populations is well suited for hypothesis generation in developing cell systems.

INTRODUCTION

Precursor cells give rise to differentiated progeny through complex developmental pathways. Single cell technologies hold the promise of elucidating the developmental progression and defining underlying transcriptomic drivers and modulators. Mass cytometry and single cell RNA sequencing (scRNAseq) can capture a high-dimensional profile of a “cellular snapshot” within analyzed tissue that contains all developing, renewing, and differentiated cell populations. High-dimensional profiles of cells can then be computationally aligned to reveal developmental relationships.

Here we show that developmental pathways can be reconstructed from single cell profiles by analyzing cells in “trajectory space,” in which each cell is represented by a profile or vector of its relative distances along nearest neighbor pathways to every other cell. The concept is illustrated in Figure 1A, with a schematic example of several cells derived from cell A and analyzed with two phenotypic markers. Cells H and E are phenotypically similar but arise from different developmental sequences and thus are developmentally distant. A dense matrix of cell-to-cell distances along the developmental pathways is constructed, which when visualized with standard dimensionality reduction tools (e.g., principal component analysis [PCA]) can be used to explore cell relationships in this novel trajectory space. As illustrated, the method reconstitutes the correct branching developmental sequences of cells in the simple example.

RESULTS AND DISCUSSION

Trajectory Space Concept and tSpace Algorithm

To implement the concept, we developed a tSpace algorithm. Its application to single cell datasets relies on the assumptions that (1) developmental processes are gradual, (2) all developmental stages are represented in the data, and (3) markers used to profile cells are regulated and sufficiently informative to distinguish different developmental pathways. Starting with cell profiles (phenotypes), tSpace identifies the K nearest neighbors of every cell, constructs a nearest neighbor (NN) graph that provides connections to all cells in the dataset, calculates distances from each cell to every other cell in the population along NN connections, and exports a dense matrix of $N \times T$ dimensions (number of cells $N \times$ number of calculated “trajectories” T , vectors of cell-to-cell distances within the manifold). Unlike other tools, tSpace determines the distances within the KNN graph using Wanderlust (Bendall et al., 2014), an algorithm that takes advantage of subgraphs and waypoints and implements a weighting scheme to reduce “short-circuits”

¹Laboratory of Immunology and Vascular Biology, Department of Pathology, School of Medicine, Stanford University, Stanford, CA 94305, USA

²The Center for Molecular Biology and Medicine, Veterans Affairs Palo Alto Health Care System and the Palo Alto Veterans Institute for Research (PAVIR), Palo Alto, CA 94304, USA

³These authors contributed equally

⁴Lead Contact

*Correspondence: ddermadi@stanford.edu (D.D.), ebutcher@stanford.edu (E.C.B.)

<https://doi.org/10.1016/j.isci.2020.100842>



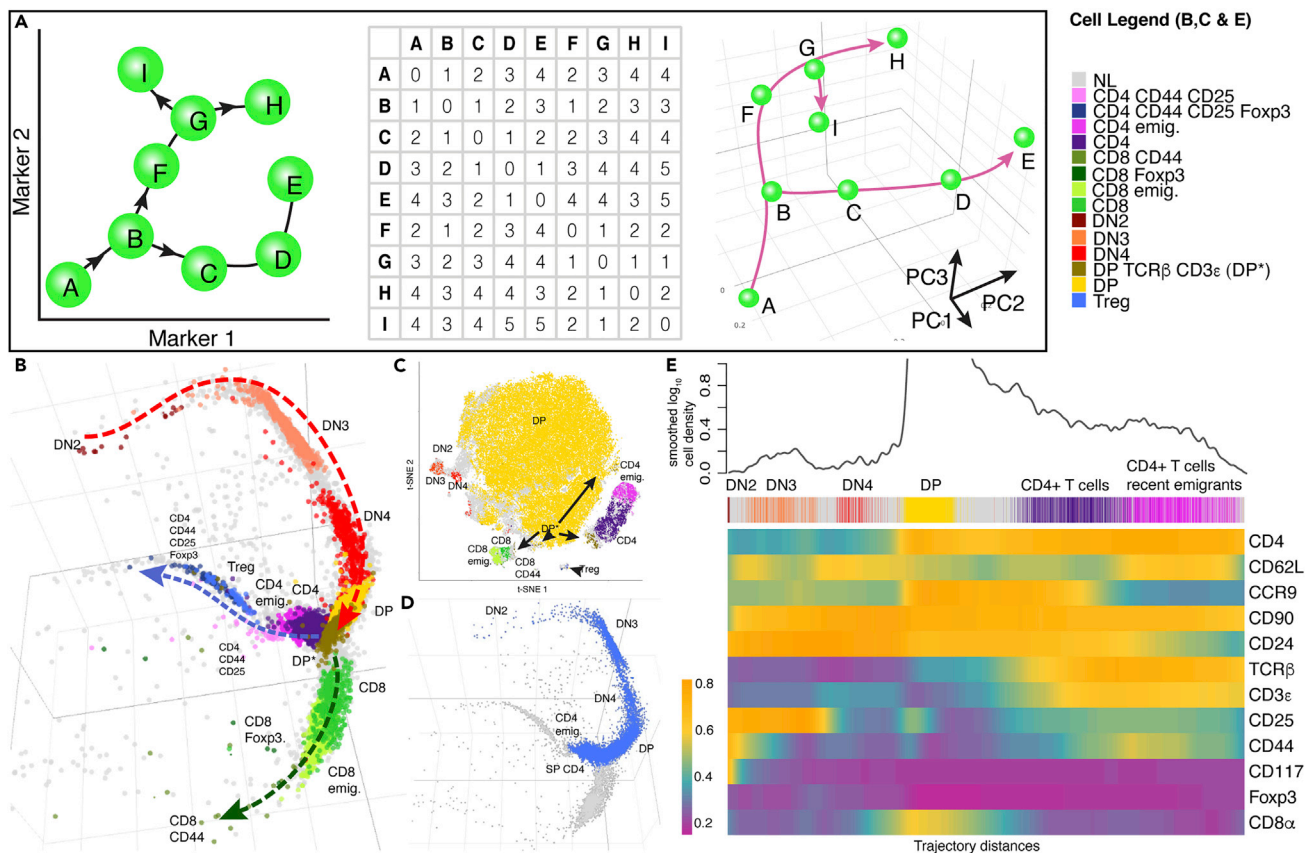


Figure 1. tSpace Concept and Application on Thymic T cells: tSpace Reveals Developmental Trajectories and Recovers Expression Patterns of Markers of T Cell Differentiation

(A) The schematic example illustrates the concept of trajectory space. The “cells” are marked with the letters A–I, and their developmental sequences are marked with arrows. Matrix of cell-to-cell distances along developmental paths is created (each cell is one unit from its nearest neighbor). Visualization of cell positions in this “trajectory space,” here using PCA, recapitulates the branches. Note that E and H, although similar phenotypically, are most distant in trajectory space, reflecting their developmental pathways.

(B) Unsupervised tSpace analysis of thymic mouse thymocytes accurately recapitulates thymic T cell development.

(C) t-SNE of thymic T cells defines clusters but not developmental relationships.

(D) Isolated trajectory from DN2 precursors to CD4 thymic emigrants.

(E) Smoothed expressions of measured markers along isolated trajectory (shown in D) reveals patterns of protein regulation during T cell differentiation. The identities of manually defined cell subsets as well as cell density along the isolated trajectory are shown for reference above the heatmap. DN, double-negative T cells; CD4 emig., CD4⁺ T cell poised emigrants; CD8 emig., CD8⁺ T cell poised emigrants.

in selecting optimal paths. The Wanderlust algorithm has been described in detail (Bendall et al., 2014). It significantly improves the definition of branching pathways even in simple flow cytometry datasets (Figure S1). We outline the effects of varying user-defined Wanderlust parameters on tSpace in the Supplemental Information (Figure S2). tSpace detection of developmental relationships is robust over a range of input parameters, allowing implementation of default settings that work well in different applications. The tSpace output provides principal component and UMAP embedding of cells in trajectory space, suitable for visualization and biological exploration of developmental pathways. The cell-to-cell distances, when exported, provide quantitative measures of phenotype change within the manifold based on user-selected metrics, useful for “pseudotime” ordering and analysis of, e.g., gene/protein expression changes along isolated linear developmental sequences. The use of unified metric-defined distances enables comparison of protein or gene expression dynamics along different trajectories on a common axis.

For samples with large cell numbers (N), tSpace has the option of calculating fewer trajectories, but it is important that these trajectories start from cells well distributed throughout phenotypical space. K-means clustering identifies groups of cells that are well distributed within phenotypic space, and we calculate

trajectories from one cell from each such cluster. The clusters are not used for further analysis. As illustration of this feature, tSpace accurately recapitulates simple branching developmental paths from as few as 25–100 trajectories (Figure S1).

To evaluate the algorithm, we applied tSpace to developing populations of lymphocytes analyzed by flow or mass cytometry, small intestine epithelial cells, and *C. elegans* analyzed by scRNAseq.

tSpace Analysis of Mouse Thymic T Cells

T cell development in the thymus is well established and allows validation of tSpace in a defined system. We generated flow cytometric profiles of mouse thymocytes using a panel of 13 antibodies (Transparent Methods). Our panel detects early T cell populations (so-called double-negative populations DN1–DN4, which lack CD4 and CD8 and are distinguished by CD44 and CD25 expression), double-positive (DP) CD4⁺CD8⁺ cells, and CD4 or CD8 single-positive (SP) T cells including poised thymic emigrant phenotype cells, regulatory T cells (CD4⁺, CD25⁺, Foxp3⁺), and a small fraction of SP T cells expressing CD44, an activation and memory marker. We manually gated on these subsets and labeled them (Figure S3) (Shah and Zuniga-Pflucker, 2014). Unsupervised tSpace analysis reveals the expected bifurcation of CD4 versus CD8 lineages from the dominant DP population in thymopoiesis and correctly positions T cells from early (DN2) to mature thymic emigrant phenotype T cells in known developmental relationships (Figure 1B). DN1 cells were not present in the dataset. In addition to the expected major bifurcation of CD4 versus CD8 cells arising from the dominant DP pool, the analysis reveals branching of regulatory T cells (Foxp3⁺) from the SP CD4 stage of CD4 branch. In contrast to methods based on or using clustering for visualization (e.g., PAGA, SPADE, p-Create, see comparisons in Supplemental Information), tSpace highlights a developmental continuum of cells allowing exploration of intermediate populations. For example, tSpace visualizes DP cells in transition to the more mature SP CD4 and CD8 T cells. The transitional cells co-express CD4 and CD8, but some have upregulated TCR β and CD3 ϵ , a characteristic of positively selected cells (Brodeur et al., 2009). Conventional clustering, based on measured markers using t-SNE, identifies the major subsets, but does not clarify developmental relationships (Figure 1C).

The tSpace output allows evaluation of expression of markers along developmental paths. To illustrate this for CD4 cell differentiation, we manually gated on cells along the path from DN2 cell to CD4 thymic poised emigrants (Figure 1D). We identified and averaged trajectories in the exported tSpace matrix (Transparent Methods) that started from early DN2 cells, and displayed marker expression along their trajectory distance from DN2 cells in a heatmap (Figure 1E). The results capture regulation of marker proteins as cells progress toward maturity, recapitulating known phenotypic progression of thymic T cell development and highlighting details of transitional states. For example, protein expression trends confirm upregulation of the chemokine receptor CCR9 in DN3 cells but reveal notably stronger expression in DN4-DP transitioning cells. CCR9 binds CCL25 expressed by thymic epithelial cells and promotes T cell cortical positioning (Wurbel et al., 2006).

tSpace Analysis of B Cell Differentiation in Tonsils and Inter-organ Trajectories with Blood

Single cell analyses hold the potential to provide insights into patterns of cell development in settings not accessible to experimental manipulation, as in the human. We applied tSpace to the development of B cells in human tonsils. Naive (IgD⁺) B cell differentiation toward immunoglobulin A or G (IgA, IgG) class switched memory or plasma cells has been investigated. However, the sequence of class switch- and fate-determining decision points and trafficking receptor induction remain poorly defined (Dufaud et al., 2017; Silva and Klein, 2015). We used a panel of mass-labeled antibodies that detects ~25 markers of B cell subsets and maturation (Transparent Methods) to stain human tonsil lymphocytes. We applied tSpace (Figure 2A) to tonsil and blood B cells and subsequently used conservative gates based on antibody staining to highlight classically defined B cell subsets for visualization in the tSpace projections (Figure S4). Cells not falling within the conservative subset gates, which include bridging populations representing transitional phenotypes, are not labeled.

tSpace analysis recapitulates developmental sequences leading from naive IgD⁺ B cells to tonsil IgG and IgA class switched effector cells. The first trajectory space principal component (tPC1) delineates the transition from naive to germinal center cells (GCC); tPC2, the differentiation of memory or plasma cells (Figures 2B and S5A); and tPC3 and tPC4 pathways, to IgA versus IgG class switched cells (Figure 2C). The distance of cells from naive B cells within the trajectory space manifold is illustrated in Figure S5B.

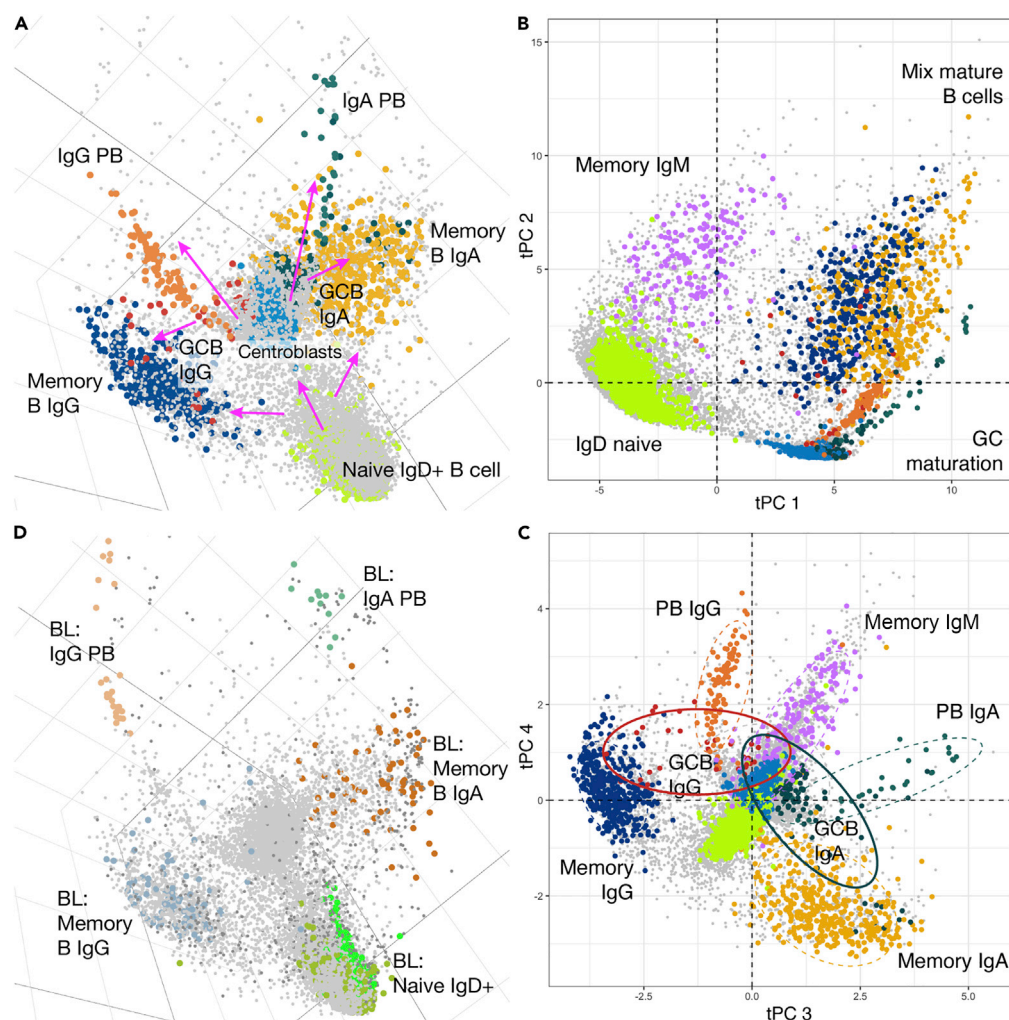


Figure 2. tSpace Analysis of B Cell Differentiation in Tonsils and Inter-organ Trajectories with Blood

(A) tSpace unravels maturation paths of B cells starting from naive B cells in tonsil throughout GC into memory B cells and plasmablasts (PB). Magenta arrows mark suggested directionalities based on known biology.

(B) tPC1 and tPC2 reveal branches and potential developmental relationships in tonsillar B cell maturation.

(C) tPC3 and tPC4 reveal branches and potential developmental relationships in tonsillar B cell maturation. Ellipses show 80% confidence intervals for indicated clusters.

(D) Blood (BL) PB align as an extension of tonsillar PB trajectories, whereas recirculating blood memory B cells overlap with the major tonsil memory cell clouds. Tonsil B cells are in light gray.

A broad strand of cells connects naive IgD⁺ B cells to proliferating germinal center (GC) centroblasts and centrocytes (Figures 2A and S5A). Along this path from naive cells, IgD is downregulated and CD77 is up-regulated as cells transition to centroblasts (Figures S5A, S5C, and S5D). There are clear, well-delineated trajectories from GCC to class switched plasmablast (PB): CD38, present on activated B cells and GCC, is further induced (Figures S5C and S5D), whereas CD20 is lost (not shown), recapitulating established patterns of antigen regulation in plasma cell development.

The regulation of trafficking receptors during B cell activation, isotype switching, and plasma cell generation in human lymphoid tissues has not been resolved. We evaluated chemoattractant and adhesion receptor expression by cells along the trajectory from naive B cells through the GC population to mature PBs. Early naive cells express CXCR5, which mediates lymphoid follicle homing, CCR6 (Figure S5C and S5D), and CCR7 (not shown). These trafficking receptors are downregulated in the transition to GCC, consistent with observations that GCC are non-migratory (Reichert et al., 1983). CD22 (Siglec2), a B cell-specific lectin that moderates B cell activation and also participates in B cell trafficking to gut-associated lymphoid tissues

(Lee et al., 2014), is maintained on GCC but lost during terminal PB differentiation (Figures S5C and S5D). Induction of IgA or IgG occurs after initial upregulation of the GC marker CD77, consistent with the known role of GC in isotype switching (Figures S5C and S5D) (Kraal et al., 1982). Homing receptors for extra-lymphoid tissue effector sites appear to be induced rapidly upon exit of cells from the GC (CD77+) pool. CCR10, a chemoattractant receptor implicated in PB migration to pulmonary and colon mucosae, is upregulated along both IgG and IgA PB lineages, whereas $\beta 7$ integrin, a component of the intestinal homing receptor, is highly upregulated in the IgA but not IgG trajectory (Figures S5C and S5D). Isotype-selective upregulation of tissue-specific adhesion receptors within a single inductive tissue has not been observed previously: the mechanisms involved may underlie the selectivity of local IgA secretion for mucosal tissues. CXCR3, implicated in lymphocyte homing to inflamed tissues (Rott et al., 1996; Seong et al., 2017), is coordinately upregulated in a minor subset of PB and by memory B cells (Figure S5A). Many tonsil memory B cells also express cutaneous lymphocyte antigen (CLA), a homing receptor for the vascular addressin CD62E associated with squamous epithelial surfaces including the oral mucosa; CLA was present on some blood PBs (not shown), but was not detected on PB branches in the tonsil.

In contrast to some other tools, tSpace does not constrain or force cells into specific developmental sequences or paths, but instead positions each cell in context with all others even when cell transitions are biologically diffuse. This is illustrated by the dispersed distribution of class switched IgG⁺ and IgA⁺ memory B cells in trajectory space (Figures 2C and S5A; and best visualized in 3D embedding, Video S1): memory cells constitute a “cloud” of cells, some of which appear to arise from the GC pool as mentioned, whereas others are closer in trajectory space to the path from naive B cells to GCs. Cell alignment in trajectory space does not intrinsically provide directional information, thus cells bridging the main memory cell population with GCs may reflect recruitment of memory cells into the active GC, or generation of memory cells from the GC reaction. The surprising alignment of many IgG- and IgA-expressing cells between naive and memory populations (adjacent to the naive to GC path; Figure S5A) suggests that, in steady-state human tonsil, activated B cells may undergo IgA or IgG class switching and conversion to memory cells without transiting through the GC reaction. Although class switch recombination is normally attributed to the GC reaction, in some mouse models class switching can occur before GC formation, and it is observed in T-independent B cell responses as well (Stavnezer and Schrader, 2014). Low expression of CD27 (Figure S5A) and retention of naive markers CCR6 and CXCR5 on the class switched cells adjacent to the “naive to GC” sequence is consistent with this interpretation (not shown). In contrast to their IgG and IgA class switched counterparts, IgM memory cells (CD27⁺, CD38⁻) are more closely connected to naive (IgM⁺, IgD⁺, CD27⁻, CD38⁻) cells in most tSpace principal components (tPCs), with tPC2 specifically expanding this trajectory (Figure 2B and Video S1). Thus, tSpace recapitulates known pathways of tonsil B cell development and differentiation, presents evidence that human B cells can follow alternative developmental paths that have only been described in animal studies, and reveals developmental stage(s) and transitions at which tissue- and inflammation-specific trafficking receptors are induced.

Developing PBs generated in lymphoid tissues leave their sites of antigen activation and circulate via the blood to distant effector sites. We reasoned that trajectories might link terminally differentiated cells, ready to exit their site of generation, with progeny cells in blood. Indeed, when we applied tSpace to combined blood and tonsil B cells datasets, blood PB aligned at the termini of tonsillar IgG and IgA PB branches (Figure 2D and Video S2). In contrast to the unidirectional path of maturing PBs, blood memory B cells and naive IgD⁺ B cells exchange between blood and lymphoid tissues through recirculation. Consistent with intermixing, these subsets overlap extensively with their tonsillar counterparts in trajectory space (Figure 2D). These results show that tSpace can unfold inter-organ transitions and elucidate developmentally programmed migration patterns of immune cells in settings wherein experimental analyses of leukocyte trafficking are challenging, as in humans.

tSpace Analysis of Mouse Small Intestine Reveals Rare Enteroendocrine Progenitor Population and Defines Transcription Factor Modules of Intestinal Differentiation

scRNAseq is emerging as a powerful tool for the characterization of cell populations and provides rich cellular profiles for studying cell relationships. We applied tSpace to published scRNAseq data from mouse intestinal epithelial cells using 2,420 variable genes (Yan et al., 2017) (Transparent Methods). Intestinal epithelium forms the single cell layer separating the lumen of small intestine from intestinal lamina propria. Almost all cells in the epithelium have a short lifespan of about 4–7 days (Barker et al., 2012), and continuous renewal is driven by division of Lgr5⁺ crypt base columnar (CBC) cells residing at the bottom of the

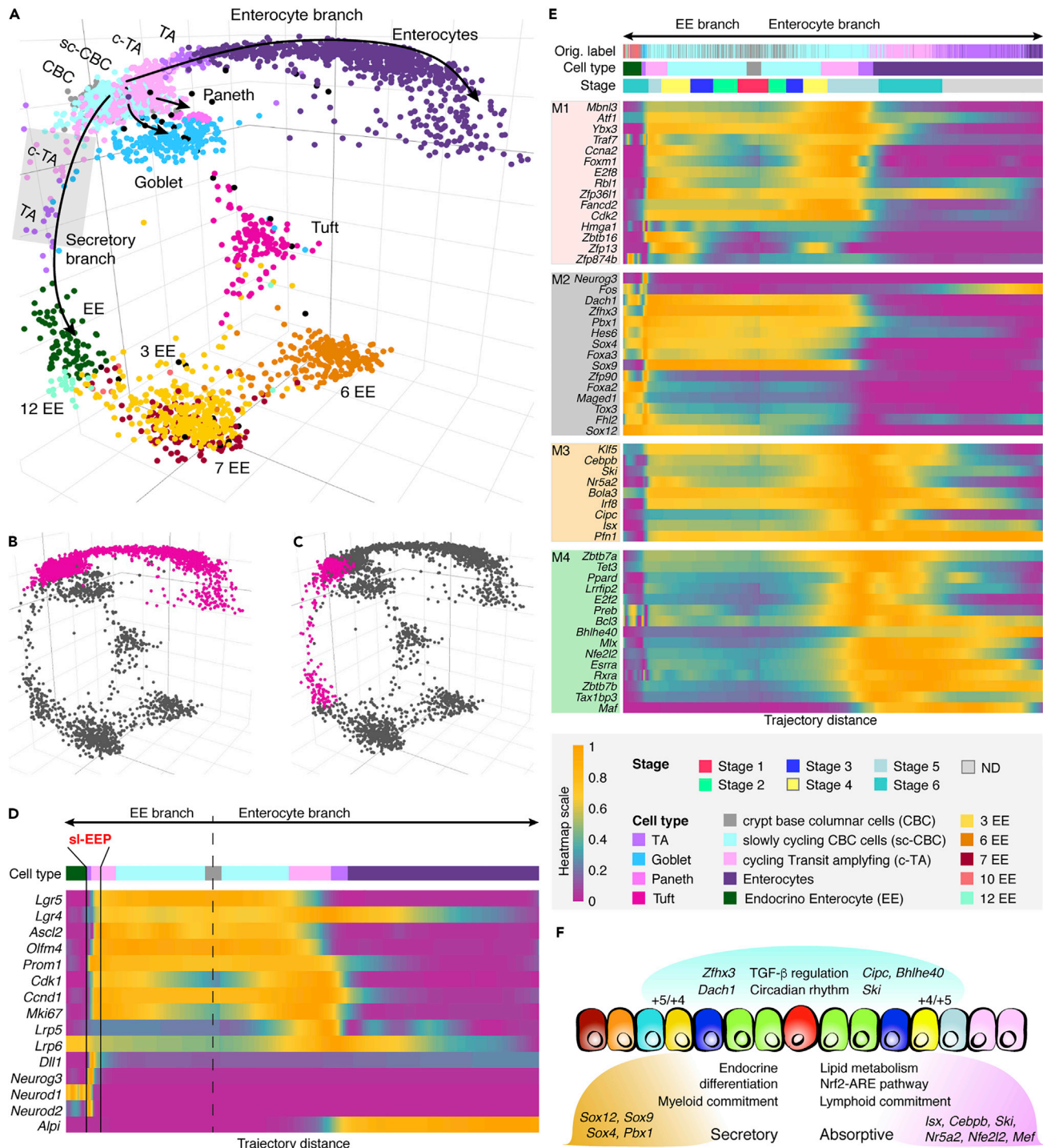


Figure 3. tSpace Analysis of Mouse Small Intestinal Epithelial Cells

(A) tSpace separates trajectories to enterocytes, enteroendocrine (EE), Paneth, and goblet cells. CBC and TA subsets were defined by our analysis, as described in Figure S7; other subsets are labeled as in Yan et al. (2017). Shaded rectangle highlights the position of short-lived EE progenitors (sIEEP) cells. (B) Isolated enterocyte trajectory. (C) Isolated EE trajectory. (D) Expression patterns of selected genes (Clevers, 2013) (known markers or regulators of intestinal crypt development; expanded gene list in Figure S8A) along the isolated trajectories.

Figure 3. Continued

(E) Four detected transcription factor modules in early trajectories, identified by comparing gene expression between cells at similar stages in the two trajectories (Transparent Methods): M1 comprises TFs involved in cell cycle and genome integrity expressed in precursor populations (early in the shared trajectory). M2 and M3-M4 differentiate the two lineages and comprise TFs that may determine cell fate or specialization (see text). Cell stage (Transparent Methods) and cell identities defined in this study (Cell type) or in Yan et al. (Orig. labels) are indicated above the heatmap. ND, fully differentiated enterocytes, not used in trajectory alignment.

(F) Summary of differences between two branches suggested by gene regulation along the trajectories. Different genes in the transforming growth factor- β and circadian rhythm pathways are expressed in the two lineages (genes in blue above the cartoon). TFs enriched in the EE branch are involved in endocrine secretory cell development, whereas TFs associated with enterocyte commitment include regulators of lipid/cholesterol metabolism. Expression of *Dll1* and *Sox4* in EE development and *Alpi* in enterocyte differentiation mark specific progenitor cells located within the +4/+5 position in the intestinal crypt according to the literature (van Es et al., 2012; Gracz et al., 2018; Tetteh et al., 2016); clear peaks are seen in their expression along the trajectories (Figures 3D and 3E) near the TA to differentiated cell transitions, likely representing these specific progenitor populations.

intestinal crypts. Transit-amplifying (TA) cells within the crypt differentiate into absorptive (enterocyte) or secretory (goblet, Paneth, tuft and enteroendocrine [EE] cell) lineages. tSpace delineates absorptive/enterocyte and secretory/EE developmental paths (Figures 3A, S6A, and S6B). Goblet and Paneth cells define short branches from the proliferating TA pool.

The ability of tSpace to position all cells in developmental relationships allows additional interesting insights. tSpace trajectories reveal a common “trunk” leading to secretory and absorptive branches, but many slow-cycling CBC (sc-CBC) and cycling TA (c-TA) cells (defined Figure S7) actually segregate to the early EE or enterocyte branches, suggesting that they are already developing toward if not committed to EE or enterocyte fates. To illustrate the application of tSpace to explore developmental progression of gene expression in this context, we isolated trajectories within the tSpace distance matrix starting from the *Lgr5*⁺ CBC cell population, gated on cells of the enterocyte branch and cells within the early segment of the EE branch preceding EE3 (Figures 3B and 3C), and plotted gene expression of cells versus their trajectory distance from CBC cells (Figure 3D). We focused initially on genes for known hallmarks of intestinal differentiation (Figures 3D and S8A) (Clevers, 2013). The analysis confirms *Ascl2* (van der Flier et al., 2009b), *Olfm4* (van der Flier et al., 2009a), and *Prom1* (Zhu et al., 2009) as robust markers of the presumptive crypt populations (CBC to TA cells) and reveals that the expression of *Prom1* extends into the TA pool, confirming previous findings (Itzkovitz et al., 2011) (Figures 3D and S8A). The Wnt agonist *Lgr5* and its homolog *Lgr4* are in resting CBC and dividing sc-CBC cells, but the analysis shows that *Lgr4* expression is retained in post-mitotic cells differentiating toward absorptive enterocytes from c-TA, suggesting that in addition to its known role in proliferation of TA cells (Mustata et al., 2011) it may contribute to enterocyte fate or specification.

Further examination revealed *Dll1*-expressing cells in trajectory space between CBC cells and mature EE populations (Figure 3A, shaded gray rectangle, Figure 3D). These cells express genes that define short-lived enteroendocrine progenitors (sLEEP) (van Es et al., 2012), which upregulate EE lineage specification genes *Neurog3*, *Neurod1*, and *Neurod2* (Figure 3D) (Jenny et al., 2002; Schonhoff et al., 2004). Consistent with their location in tSpace projection, sLEEP are well-documented precursors of EE cells (Barker et al., 2012; van Es et al., 2012; Schonhoff et al., 2004). The EE branch proceeds through EE3 cells, recently identified as EE intermediates, giving rise to specialized mature EE subsets (Yan et al., 2017) (for cell labels see Transparent Methods and Figure S7). Interestingly, sparse intermediates link a single tuft cell population to both CBC/TA and to EE3 cells (best visualized when UMAP is applied to trajectory space matrix, Figures S6A and S6B). Although the number of intermediate cells linking these two pathways to tuft cells would suggest caution in interpretation, it is noteworthy that a dual origin of tuft cells (directly from *Lgr5*⁺ CBC cells but also from EE3 EE cells) has been proposed from multiple lines of evidence (Gerbe et al., 2012; Yan et al., 2017). tSpace performed well when compared with SPADE, a minimum spanning tree (MST) algorithm applied to visualize trajectory relationships in the original analysis of this scRNAseq dataset. SPADE (Yan et al., 2017) and tSpace both delineate the major CBC to enterocyte and EE branches, the relationship of goblet and Paneth cells to CBC/TA, and the terminal branching of EE subsets. SPADE generates a 2D representation of an MST structure, an approach that is inherently challenged by non-tree-like developmental paths, such as the paths from EE3 and CBC that converge on tuft cells (Figures 3A, S6A, and S6B). tSpace identified a single tuft cell pool with dual connections, whereas SPADE analysis forced tuft cells into two disconnected populations, one arising from CBC cells and the other from intermediates leading to EE3 cells. SPADE also failed to detect or properly position sLEEP on the path to EE cells (Yan et al., 2017). In contrast to tSpace, which positions each cell in trajectory space, SPADE and related MST algorithms rely on prior definition of cell clusters and limited gene sets, features that run the risk of missing or mislabeling

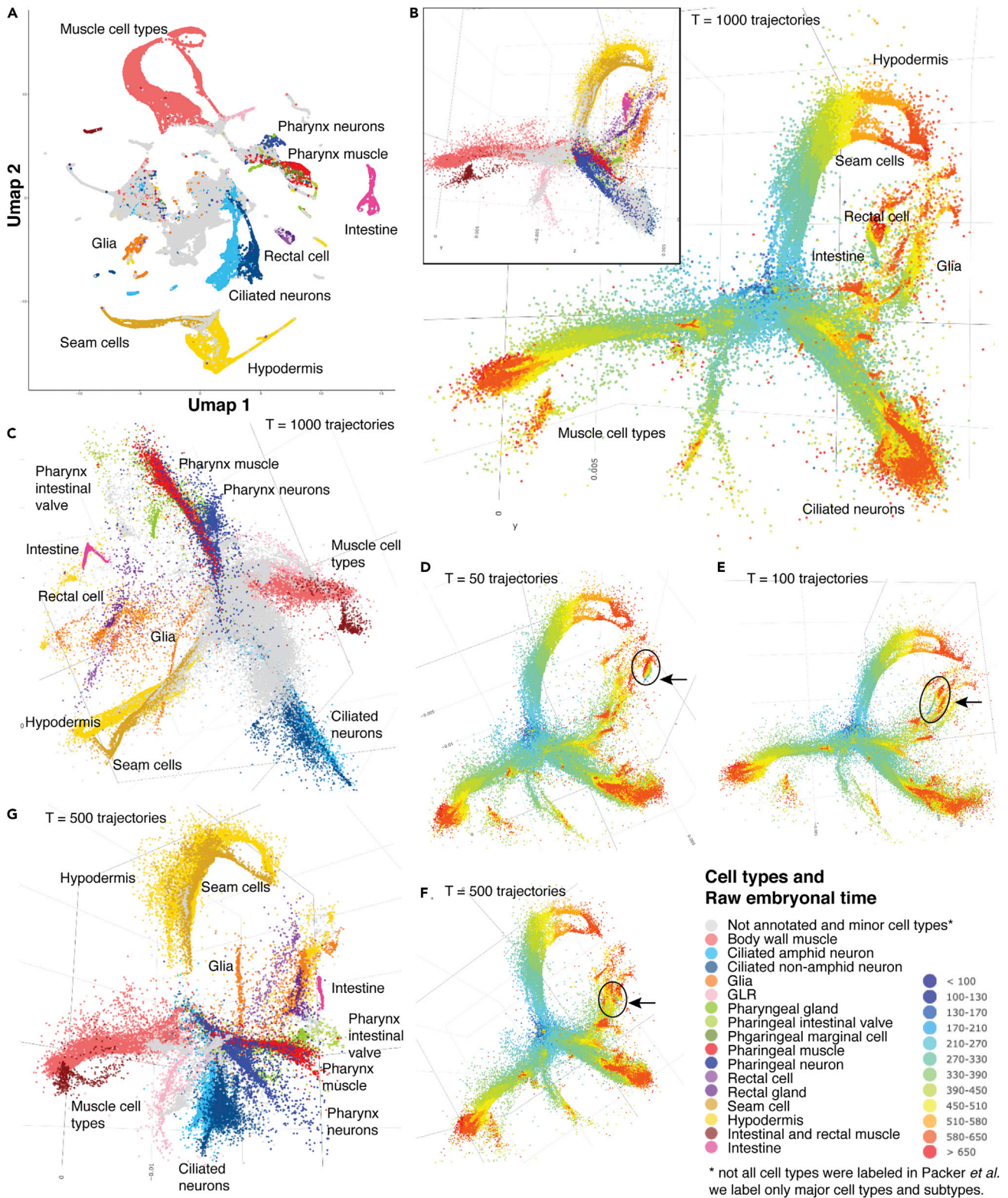


Figure 4. Benchmarking of tSpace Performance Using a Developmentally Timed *C. elegans* Dataset

(A) UMAP of 75 PCs of gene expression matrix, re-creating the UMAP from Packer *et al.*, showing major cell types.

(B) tSpace analysis using 1,000 trajectories, showing the first three tPCs and raw embryonic time; inset shows cell types.

Figure 4. Continued

(C–F) (C) The same tSpace analysis as in (B) visualized in tPC1, tPC3, and tPC4: this combination of tPCs allows separation of all major lineages, and interestingly places pharyngeal intestinal valve (PIV) cells in close proximity to intestine and intestine close to rectal cells, relations that resemble spatial cell positions in *C. elegans* body: PIV cells link the posterior bulb of the pharynx to the anterior cells of the intestine, and the intestine is eventually connected to the rectum. The first three tPCs and raw embryonal time using (D) 50 trajectories, (E) 100 trajectories, and (F) 500 trajectories. Arrows and ellipsoids mark intestinal cell type.

(G) tSpace analysis using 500 trajectories, 3D representation of cell types in tPC1, tPC3, and tPC4. tSpace defines branches for all main lineages.

important cell intermediates (Yan et al., 2017). sLEEP were defined either as cycling CBC or goblet cells in the published analysis and were subsumed in biologically inappropriate clusters (Figures S7A, S7B, S7D, and S8A see original labels).

Overall, cell positioning in trajectory space and the patterns of gene expression reflect observations from decades of research on intestinal development, and also suggest refinements to current understanding. Many TA cells express gene programs leading to secretory versus absorptive phenotypes (Figures 3D and S8A), indicating that fate selection is already initiated within the dividing (TA) pool that arises from Lgr5⁺ CBC. A global survey of transcription factor (TF) expression during early specification of secretory versus absorptive fates has not been described. We evaluated gene expression along tSpace-defined developmental sequences to identify TF that might specify and/or control downstream cell specialization. Four distinct TF modules were identified (Transparent Methods) based on their patterns of regulation along early EE or enterocyte branches (M1–M4, Figures 3E and S8C). Genes for proliferation and DNA maintenance (M1, e.g., *Ccna2*, *Cdk2*, *Fancd2*, *Rbl1*) are expressed by dividing sc-CBC and TA “early” along the trajectory, as expected. A second module of TF genes is also expressed by early cells but is maintained selectively in the EE branch: these include TFs associated with endocrine and pancreatic development (e.g., *Foxa2*, *Foxa3*, *Neurog3*, *Sox4*, *Sox9*) that may coordinate secretory pathways within intestinal EE cells (Tabula Muris Consortium et al., 2018). Interestingly, among these, tSpace revealed an unexpectedly high and selective expression of *Sox4* in sLEEP cells, suggesting it as a novel candidate contributor to EE specification (Figures 3E, S7C, and S7E): this prediction has been subsequently confirmed (Gracz et al., 2018). Module 3 and 4 TFs are expressed preferentially in the enterocyte branch. They include TF involved in lipid and cholesterol metabolism required for mature enterocytes (e.g., *Cebpb*, *Klf5*, *Nr5a2*, Figures 3E and S8C) (Degirolamo et al., 2015; Yen et al., 2014), and also *Nfe2l2* and *Maf* associated with the activation of *Nfe2l2*/*Nrf2*-antioxidant response element (ARE) pathway (Itoh et al., 1997). Enterocytes utilize short fatty acids as a source of energy; fatty acid metabolism generates reactive oxygen species (ROS), and ROS are also abundant in the intestinal lumen. Upregulation of the *Nfe2l2*-ARE pathway may help protect differentiating enterocytes from oxidative damage (Ferrebee et al., 2018). The analysis also identified *Isx* and *Ski* (both within M3) as putative novel markers of TA cells within the early enterocyte developmental branch; lack of specific markers has hindered isolation of TA cells and further probing of their plasticity.

Benchmarking of tSpace Performance Using a Developmentally Timed *C. elegans* Dataset

Finally, we benchmarked tSpace using *C. elegans* scRNAseq (Packer et al., 2019), a dataset that can be considered the gold standard for developmental trajectory inference benchmarking because each cell is associated with actual developmental time. We run tSpace analysis using all 86,024 cells and 75 PCs in gene expression space (Transparent Methods). tSpace provides connections to all cell types, unlike UMAP, the method used by Packer et al. (Figures 4A and 4B). Using UMAP on gene profiles looks similar, as expected, to the UMAP in Packer et al. (2019) Major cell populations are well separated from each other without indication of developmental relations: for example, groups of muscle and pharynx cell types are disconnected from the rest of the cells, as well as intestine or hypodermis and seam cells. Conversely, tSpace analysis using 1,000 trajectories aligns all cells in the correct order as examined by plotting raw embryonal time associated with every cell in the dataset (Figure 4B) and separates all major cell lineages (Figures 4B, inset, and 4C). tSpace alignments using 50 trajectories still retained distinct branching of all major cellular lineages and the correct developmental connections between the cells (Figures 4D–4G). As few as 50 trajectories permits tSpace to order cells correctly (Figure 4D), but alignments using 100 and 500 trajectories showed progressive improvement in resolution and connection between cell types (e.g., compare intestinal cells marked with an arrow Figures 4D–4F). Thus, although tSpace ran into memory limitations when the trajectory number *T* is high (e.g., 5,000 trajectories failed when applied to the current example), *T* = 100 is sufficient for even large and complex datasets.

Similar to the UMAP analysis (Packer et al., 2019), we noticed that tSpace provides greater resolution toward terminally differentiated cell types. Packer et al. iteratively run UMAP on subsets of cell types to

achieve resolution and precision in the early developmental stages or specific tissues. To examine tSpace performance in early development, we focused on ABpxppp sublineage (Figure S9A, for details see Packer et al., 2019). Indeed, utilizing UMAP on a full dataset does not allow precise mapping and does not preserve developmental connections of the early cell lineages (Figures S9B and S9C). Remarkably, tSpace connects and correctly orders ABpxppp sublineage (Figures S9D–S9I). Although the cells are correctly ordered, unlike in UMAP, the two cycles of division after ABpxppp are difficult to visually dissect. However, seven daughter cell types (ABpxpppaaa, ABpxpppaap, ABpxpppapa, ABpxppppaa, ABpxppppap, ABpxpppppa, ABpxpppppp), products of the third division after ABpxppp, are separated and easily recognized as elongated branches (Figures S9E–S9H). tSpace maps one of them, ABpxpppppa, close to, but not as part of the muscle cell-type branch (Figure S9I) reflecting the true biology. Muscle cell types are developed from mesoderm MS lineage. UMAP of the full dataset incorrectly suggests ABpxpppppa as a precursor of the MS lineage and muscle cells. Overall, tSpace preserves developmental cell-to-cell relations better than UMAP and performs well in real-world complex data with 5+ diverse lineages (MS muscle, MS pharynx, C and D muscle, and AB-derived lineages) represented by *C. elegans*.

tSpace Performance and Comparison with Other Trajectory Inference Algorithms

tSpace applied to four different datasets illustrates how the number of trajectories, of cells, and of measured protein or gene parameters all impact computation time (Table S1), as does the hardware employed. For cytometry datasets, we generally run tSpace on all measured parameters. For scRNAseq studies, we often run the algorithm with significant principal components. For any larger data, especially in number of cells, we suggest use of multicore computers to speed up the analysis. As the examples presented illustrate, we generally limit T to 100–1,000. Although not illustrated here, for scRNAseq and other sparse datasets we find that imputation can further improve the definition of branches and branchpoints along developmental trajectories.

The tSpace approach is conceptually similar to isomap (Tenenbaum et al., 2000). Both methods provide a global approach to dimensionality reduction, designed to preserve manifold geometry at all scales. Both algorithms determine geodesic distances along a KNN graph. Isomap embeds the resulting distance matrix in low dimensions using multidimensional scaling (MDS). It has been successfully applied to diverse high-dimensional datasets (Hannachi and Turner, 2013; Mahecha et al., 2007; Silva and Tenenbaum, 2003; Stamati et al., 2010), but it has not been widely adopted for high-dimensional single cell analyses, perhaps because of well-described limitations. The algorithm is computationally and memory expensive. This has been addressed in part in “landmark isomap” by calculating approximate distances using a set of randomly selected “landmark” cells. To ensure uniform sampling of the manifold, we modify this approach in tSpace by selecting individual cells from each of T K-means clusters, where T is the number of trajectories to be calculated. We show that linear trajectories (distance vectors) calculated from 100–250 well-distributed starting cells are sufficient to recapitulate cellular relationships in each of the datasets here. As each such trajectory includes all cells and provides estimates of their intercellular distances within the trajectory, a small number of such trajectories provides a surprisingly good approximation of cell relationships in trajectory space. Isomap suffers also from sensitivity to “short circuit” errors if K is too large or if noise in the data positions cells aberrantly between valid branches or populations in the manifold. Short circuits pose a problem with the Dijkstra algorithm, used in isomap to calculate shortest paths between cells. We take advantage of Wanderlust, which refines distances and avoids “short circuits” by using subgraph averaging and weighting in shortest path calculations based on waypoints (Bendall et al., 2014). We show that tSpace with Wanderlust improves the definition of developmental paths (Figure S1). However, isomap uses MDS, a memory-intensive algorithm, for dimensionality reduction and visualization of manifold relationships, tSpace utilizes PCA and/or UMAP. We find that the first three tPCs often embody the most important developmental branches (with simple branching development), but higher tPCs can also reveal critical biological processes. Many methods allow reconstruction of simple developmental branching sequences, but it is becoming increasingly clear that differentiating cells in development and cancer can and often do retain multi-potency even as they mature. This leads to complex higher-dimensional “lineages” or developmental pathways that cannot be represented in 2D or even 3D. In this setting, exploration of PCA projections of trajectory space is valuable, as higher tPCs can reveal additional branching pathways and relationships. We illustrate, for example, the parallel pathways of IgA versus IgG memory and plasma cell development in tonsil B cells, which dominate the fourth tPC. tSpace also implements dimensionality reduction with UMAP (McInnes et al., 2018). Although UMAP implements a force-directed algorithm that obscures developmental distances, it is an excellent tool for reducing dimensionality to

2-3 dimensions for visualization of the tSpace manifold. We show for our scRNAseq example that UMAP embedding of trajectory space reveals developmental branches better than UMAP embedding of the original gene expression matrix (see [Supplemental Information: Comparison of tSpace and Other Trajectory Inference Algorithms](#)).

We examine the performance of several widely used methods (Reviewed in [Cannoodt et al., 2016](#); [Saelens et al., 2018](#)) with tSpace in [Supplemental Information \(Figures S10–S14\)](#). Few unsupervised methods exist (we tested Monocle, p-creode, DPT, PAGA, and because of its speed and usefulness UMAP), and none combine the scalability and quantitative distance metrics of tSpace. For example, tSpace has advantages over algorithms that use memory-intensive MST methods to define branch points, as for large datasets these depend on downsampling of cells or calculation of relationships between clusters (rather than individual cells) to reduce computational complexity. Examples include slingshot ([Street et al., 2018](#)), p-creode ([Herring et al., 2017](#)), and SPADE ([Qiu et al., 2011](#)). As highlighted above in the discussion of published analysis of intestinal epithelial cells, downsampling in MST-based methods holds inherent risks of obscuring important cell subsets, and in most algorithms, fails to position each cell in developmental relationships. tSpace avoids the loss of individual cell resolution associated with cell downsampling, while retaining the ability to reveal the developmental relationships of all cells to each other even when popular and faster UMAP fails (e.g., analysis of full *C. elegans* dataset). Algorithms that focus on computationally defining branchpoints and tree structures (e.g., Monocle, slingshot) can also limit appreciation of alternative pathways of differentiation represented by cells that bridge between dominant pathways (e.g., converging paths of tuft cells). Moreover, in contrast to algorithms that rely only on local cell relationships or that use force-directed graph methods, the global approach of tSpace estimates distant as well as nearby cell relationships within the manifold. Indeed, the algorithm exports a dense matrix of meaningful cell-to-cell distances that represent measures of the extent of phenotypic change along developmental pathways. As illustrated in our examples, cells along specific developmental pathways and branches can be easily gated (isolated) in plots of tPCs using commonly available software such as Flowjo, JMP, or in R ([Transparent Methods](#)). Trajectories starting from branch termini or other desired points within pathways are readily identified within the tSpace matrix and plotted versus gene/protein expression to characterize changes in cell phenotypes along isolated developmental sequences (as in [Figures 1, 2, and 3](#)).

Conclusions

We have presented the concept of trajectory space and its implementation in the tSpace algorithm for elucidation of branching or convergent developmental pathways and mechanisms from single cell profiles. tSpace performs well across different biological systems and platforms, and reveals known and novel biology. tSpace embodies a combination of useful features including (1) applicability to any type of data (proteomic, transcriptomic, etc.), (2) simplicity of use, (3) robustness to input parameters, (4) scalability and independence from the need for cell downsampling, (5) positioning of each individual cell correctly in developmental relationships (allowing visualization of alternative or minor pathways of differentiation), (6) retention of global as well as local cell relationships with export of quantitative measures of cell-to-cell distances in the manifold, and (7) independence from requirements of clustering or prior information. The tSpace outputs are often more precise than other trajectory inference algorithms, reproducible, intuitive, and amenable to exploration of biology (gene or protein expression, trajectory isolation, etc.). We believe that tSpace will prove useful to the rapidly growing field of single cell analysis.

Limitations of the Study

We highlight some of the practical features and limitations of tSpace and of other published algorithms designed for multi-branching and complex manifolds ([Table S2](#)). tSpace attempts a true representation of the underlying manifold, without manipulation for graphical expedience, and when visualized in principal component projection is highly reproducible and robust to changes of parameters: it allows users to explore hidden biology and compare results confidently. However, faithfulness to the manifold limits the ability of tSpace, in comparison to user-optimized UMAP for example, to generate appealing 2D or 3D representations. We discuss UMAP reduction of the tSpace matrix as a compromise solution. Similarly, whereas UMAP implements features that result in attraction and repulsion to enforce cluster formation and separation, tSpace does not. Because of this tSpace excels at revealing sparsely populated developmental trajectories that may be 'broken' in cluster-forming methods. However, tSpace can make links between unrelated cell types when biologically valid intermediates are missing, and as in other methods, computationally defined trajectories must be interpreted in the light of prior biological knowledge and/or be validated

experimentally. tSpace faces memory limits in cases of large datasets (100,000+ cells) when calculating several thousand trajectories; and tSpace running times in such instances can be a limiting factor. In practice, we find that $T = 100$ provides accurate trajectories for most datasets, and tSpace in Matlab with $T = 100$, 30,000 cells and 30 principal components from scRNAseq data runs on a personal computer in under 20 minutes. A final limitation of this study is that we lack a quantitative metric to compare the usefulness of tSpace (vs other methods considered) for manifold exploration and biological interpretation, relying instead on concrete though anecdotal examples.

METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

DATA AND CODE AVAILABILITY

This manuscript is accompanied with Supplemental items. Code and data are available upon request.

Previously published dataset of intestinal cell populations was provided by the authors as normalized and scaled expression values (Yan et al., 2017). *C. elegans* dataset (86,024 cells) was downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE126954.

tSpace package for R is available on <https://github.com/hylasD/tSpace>, and MATLAB https://github.com/hylasD/MATLAB_version_tSpace. tSpace tutorial can be found at <http://denisdermadi.com/tSpace-trajectory-inference-algorithm>.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.100842>.

ACKNOWLEDGMENTS

Mass cytometry analysis for this project was done on Cyran instrument in the Stanford Shared FACS Facility, obtained by S10OD016318-01 NIH grant. This study is the result of work supported with resources and the use of facilities at the VA Palo Alto Health Care System (Disclaimer: the contents of this study do not represent the views of VA or the United States Government). We thank Raghav and Durga Ganesh for assistance with parts of the code, Menglan Xiang and Sofia Nordling for help with evaluations of the algorithm, Steven Schaffert for constructive discussions, and Lourdes Magalhaes for administrative assistance.

This work was supported by NIH grants R37-AI047822, R01 AI130471, and R01-CA228019 and award I01 BX-002919 from the Dept of Veterans Affairs to E.C.B.; by R01-AI109452 to H.H., and by pilot awards under ITI Seed 122C158 and CCSB grant U54-CA209971 to D.D. and E.C.B. M.B. was supported by fellowships from the German Research Foundation (DFG, BS56/1-1) and the Crohn's and Colitis Foundation of America. A.S. was supported by the Mobility Plus fellowship from the Ministry of Science and Higher Education, Poland (1319/MOB/IV/2015/0).

AUTHOR CONTRIBUTIONS

D.D. wrote the algorithm, supervised computational analyses, and interpreted intestinal and *C. elegans* data; K.B. wrote parts of the algorithm; D.D. and H.H. designed and interpreted the thymus study; M.B. and N.H.L. performed and D.D. and M.B. analyzed the tonsil B cell study; A.S. prepared human tissues; D.D., M.B., and E.C.B. wrote the manuscript; H.H. provided advice; E.C.B. conceived the trajectory space concept and supervised the project; D.D. and E.C.B. revised manuscript.

DECLARATION OF INTERESTS

The authors have declared no conflict of interest.

Received: July 10, 2019

Revised: December 17, 2019

Accepted: January 10, 2020

Published: February 21, 2020

REFERENCES

- Barker, N., van Oudenaarden, A., and Clevers, H. (2012). Identifying the stem cell of the intestinal crypt: strategies and pitfalls. *Cell Stem Cell* 11, 452–460.
- Bendall, S.C., Davis, K.L., Amir, E., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157, 714–725.
- Brodeur, J.-F., Li, S., Damlaj, O., and Dave, V.P. (2009). Expression of fully assembled TCR–CD3 complex on double positive thymocytes: synergistic role for the PRS and ER retention motifs in the intra-cytoplasmic tail of CD3ε. *Int. Immunol.* 21, 1317–1327.
- Cannoodt, R., Saelens, W., and Saeys, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.* 46, 2496–2506.
- Clevers, H. (2013). The intestinal crypt, A prototype stem cell compartment. *Cell* 154, 274–284.
- Degiralamo, C., Sabbà, C., and Moschetta, A. (2015). Intestinal nuclear receptors in HDL cholesterol metabolism. *J. Lipid Res.* 56, 1262–1270.
- Dufaud, C.R., McHeyzer-Williams, L.J., and McHeyzer-Williams, M.G. (2017). Deconstructing the germinal center, one cell at a time. *Curr. Opin. Immunol.* 45, 112–118.
- Ferrebee, C.B., Li, J., Haywood, J., Pachura, K., Robinson, B.S., Hinrichs, B.H., Jones, R.M., Rao, A., and Dawson, P.A. (2018). Organic solute transporter α - β protects ileal enterocytes from bile acid-induced injury. *Cell Mol. Gastroenterol. Hepatol.* 5, 499–522.
- Gerbe, F., Legraverend, C., and Jay, P. (2012). The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* 69, 2907–2917.
- Gracz, A.D., Samsa, L., Fordham, M.J., Trotier, D.C., Zwarycz, B., Lo, Y.-H., Bao, K., Stamer, J., Raab, J.R., Shroyer, N.F., et al. (2018). SOX4 promotes ATOH1-independent intestinal secretory differentiation toward tuft and enteroendocrine fates. *Gastroenterology* 155, 1508–1523.e10.
- Hannachi, A., and Turner, A.G. (2013). Isomap nonlinear dimensionality reduction and bimodality of Asian monsoon convection. *Geophys. Res. Lett.* 40, 1653–1658.
- Herring, C.A., Banerjee, A., McKinley, E.T., Simmons, A.J., Ping, J., Roland, J.T., Franklin, J.L., Liu, Q., Gerdes, M.J., Coffey, R.J., and Lau, K.S. (2017). Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.* 6, 37–51.e9.
- Itoh, K., Chiba, T., Takahashi, S., Ishii, T., Igarashi, K., Katoh, Y., Oyake, T., Hayashi, N., Satoh, K., Hatayama, I., et al. (1997). An Nrf2/small Maf heterodimer mediates the induction of phase II detoxifying enzyme genes through antioxidant response elements. *Biochem. Biophys. Res. Commun.* 236, 313–322.
- Itzkowitz, S., Lyubimova, A., Blat, I.C., Maynard, M., van Es, J., Lees, J., Jacks, T., Clevers, H., and van Oudenaarden, A. (2011). Single-molecule transcript counting of stem-cell markers in the mouse intestine. *Nat. Cell Biol.* 14, 106–114.
- Jenny, M., Uhl, C., Roche, C., Duluc, I., Guillermin, V., Guillemot, F., Jensen, J., Kedinger, M., and Gradwohl, G. (2002). Neurogenin3 is differentially required for endocrine cell fate specification in the intestinal and gastric epithelium. *Embo J.* 21, 6338–6347.
- Kraal, G., Weissman, I., and Butcher, E. (1982). Germinal centre B cells: antigen specificity and changes in heavy chain class expression. *Nature* 298, 377–379.
- Lee, M., Kiefel, H., LaJevic, M.D., Macauley, M.S., Kawashima, H., O'Hara, E., Pan, J., Paulson, J.C., and Butcher, E.C. (2014). Transcriptional programs of lymphoid tissue capillary and high endothelium reveal control mechanisms for lymphocyte homing. *Nat. Immunol.* 15, 982–995.
- Mahecha, M.D., Martínez, A., Lischeid, G., and Beck, E. (2007). Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecol. Inform.* 2, 138–149.
- McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv. <https://arxiv.org/abs/1802.03426v2>.
- Mustata, R.C., Loy, T., Lefort, A., Libert, F., Strollo, S., Vassart, G., and Garcia, M. (2011). Lgr4 is required for Paneth cell differentiation and maintenance of intestinal stem cells ex vivo. *EMBO Rep.* 12, 558–564.
- Packer, J.S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., et al. (2019). A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* 365, eaax1971.
- Qiu, P., Simonds, E.F., Bendall, S.C., Jr., 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. *Nat. Biotechnol.* 29, 886.
- Reichert, R., Gallatin, W., Weissman, I., and Butcher, E. (1983). Germinal center B cells lack homing receptors necessary for normal lymphocyte recirculation. *J. Exp. Med.* 157, 813–827.
- Rott, L., Briskin, M., Andrew, D., Berg, E., and Butcher, E. (1996). A fundamental subdivision of circulating lymphocytes defined by adhesion to mucosal addressin cell adhesion molecule-1. Comparison with vascular cell adhesion molecule-1 and correlation with beta 7 integrins and memory differentiation. *J. Immunol.* 156, 3727–3736.
- Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *Biorxiv*, 276907, <https://doi.org/10.1101/276907>.
- Schonhoff, S.E., Giel-Moloney, M., and Leiter, A.B. (2004). Neurogenin 3-expressing progenitor cells in the gastrointestinal tract differentiate into both endocrine and non-endocrine cell types. *Dev. Biol.* 270, 443–454.
- Seong, Y., Lazarus, N.H., Sutherland, L., Habtezion, A., Abramson, T., He, X.-S., Greenberg, H.B., and Butcher, E.C. (2017). Trafficking receptor signatures define blood plasmablasts responding to tissue-specific immune challenge. *JCI Insight* 2, e90233.
- Shah, D., and Zuniga-Pflucker, J. (2014). An overview of the intrathymic intricacies of T cell development. *J. Immunol.* 192, 4017–4023.
- Silva, N.S., and Klein, U. (2015). Dynamics of B cells in germinal centres. *Nat. Rev. Immunol.* 15, 137–148.
- Silva, V.D., and Tenenbaum, J.B. (2003). Global versus Local Methods in Nonlinear Dimensionality Reduction, pp. 721–728.
- Stamati, H., Clementi, C., and Kavraki, L.E. (2010). Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins* 78, 223–235.
- Stavnezer, J., and Schrader, C.E. (2014). IgH chain class switch recombination: mechanism and regulation. *J. Immunol.* 193, 5370–5378.
- Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* 19, 477.
- Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372.
- Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319–2323.
- Tetteh, P.W., Basak, O., Farin, H.F., Wiebrands, K., Kretschmar, K., Begthel, H., van den Born, M., Korving, J., de Sauvage, F., van Es, J.H., et al. (2016). Replacement of lost lgr5-positive stem cells through plasticity of their enterocyte-lineage daughters. *Cell Stem Cell* 18, 203–213.
- van der Flier, L.G., Haegebarth, A., Stange, D.E., van de Wetering, M., and Clevers, H. (2009a). OLFM4 is a robust marker for stem cells in human intestine and marks a subset of colorectal cancer cells. *Gastroenterology* 137, 15–17.
- van der Flier, L.G., van Gijn, M.E., Hatzis, P., Kujala, P., Haegebarth, A., Stange, D.E., Begthel, H., van den Born, M., Gurjev, V., Oving, I., et al. (2009b). Transcription factor Achaete scute-like 2 controls intestinal stem cell fate. *Cell* 136, 903–912.
- van Es, J.H., Sato, T., van de Wetering, M., Lyubimova, A., Nee, A., Gregorieff, A., Sasaki, N.,

Zeinstra, L., van den Born, M., Korving, J., et al. (2012). Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. *Nat. Cell Biol.* *14*, 1099.

Wurbel, M., Malissen, B., and Campbell, J.J. (2006). Complex regulation of CCR9 at multiple discrete stages of T cell development. *Eur. J. Immunol.* *36*, 73–81.

Yan, K.S., Gevaert, O., Zheng, G., Anchang, B., Probert, C.S., Larkin, K.A., Davies, P.S., Cheng, Z., Kaddis, J.S., Han, A., et al. (2017). Intestinal enteroendocrine lineage cells possess homeostatic and injury-inducible stem cell activity. *Cell Stem Cell* *21*, 78–90.e6.

Yen, C.-L., Nelson, D.W., and Yen, M.-I. (2014). Intestinal triacylglycerol synthesis in fat absorption

and systemic energy metabolism. *J. Lipid Res.* *56*, <https://doi.org/10.1194/jlr.r052902>.

Zhu, L., Gibson, P., Currie, S.D., Tong, Y., Richardson, R.J., Bayazitov, I.T., Poppleton, H., Zakharenko, S., Ellison, D.W., and Gilbertson, R.J. (2009). Prominin 1 marks intestinal stem cells that are susceptible to neoplastic transformation. *Nature* *457*, 603–607.

iScience, Volume 23

Supplemental Information

Exploration of Cell Development Pathways

through High-Dimensional Single Cell

Analysis in Trajectory Space

Denis Dermadi, Michael Bscheider, Kristina Bjegovic, Nicole H. Lazarus, Agata Szade, Husein Hadeiba, and Eugene C. Butcher

Supplemental materials

Formal description of the tSpace algorithm

Every analyzed cell c_i is defined by a vector of expression values for each measured marker (v) $c_i = (v_1, v_2, v_3, \dots, v_j)$, where i is between 1 and the total number of cells (N) and j is the number of measured markers (m). Profiles of all cells within a sample form an N -by- m input matrix. Prior to application of tSpace, the input matrix has to be cleaned of outliers and artifacts and adequately transformed, e.g. logicle for FACS data, the inverse hyperbolic sine function for CyTOF and logarithmic for scRNAseq.

In the *trajectory space* concept, each cell is represented as a vector of distances to every other cell in the data set. Implementation of the concept through the tSpace algorithm involves three steps.

In the first step tSpace calculates a K-nearest neighbor (KNN) graph, in which there is an edge from cell c_i to c_j if and only if the distance between c_i and c_j is one of the K smallest distances between c_i and all other cells. The user-defined neighborhood size (K), preferably small, is selected to allow inclusion of every cell in the graph. Values of $K \sim 20$ -30 works for most data sets in our analysis. In order to reduce “short-circuits”, which may occur in basic KNN graph, L ($L < K$) number of connections for every cell is preserved and a L-KNN sub-graph is created. User defines how many sub-graphs will be calculated by defining a parameter G .

The second step of tSpace is computation of a *trajectory space* distance matrix. This step utilizes parallelization for performance improvement. tSpace computes distances [$d_{Wanderlust}(x, y)$] from each cell to every other cell within the L-KNN graph using a modified Dijkstra algorithm, originally published under the name Wanderlust (Bendall et al., 2014). Briefly, for each cell c_i , Wanderlust initially calculates interim trajectories for each of G sub-graphs. These interim trajectories are averaged to generate a consensus distance to every cell. Cell distances within each sub-graph are calculated in relationship to both the start cell and to “waypoint cells” that are randomly picked from within uniformly distributed subsections of an initial trajectory. The number of waypoints (WP) is user defined. Distances from the start cell to every other cell are iteratively refined using the waypoints and an exponential weighting scheme until convergence (reaching correlation of 0.9999 between the trajectories). The weight matrix is determined using the distances between waypoints and the rest of the cells; this improves calculated distances by giving more importance to the distances around neighboring waypoints. The final *trajectory space* matrix M is a dense matrix of refined distances (d) from each cell to every cell in the dataset, where $M_{ij} = d_{Wanderlust}(c_i, c_j)$, $i \& j \in [1, N]$, with N being the total number of measured cells, and each cell is described by a vector of distances $c_i = (d_1, d_2, d_3, \dots, d_j)$, $i \& j \in [1, N]$. We refer to the vector of distances from cell c_i to every other cell as a trajectory.

Lastly, dimensionality reduction (e.g. PCA or UMAP) is used to visualize cells and their relationships in trajectory space.

In the ideal situation, the number of trajectories T is equal to the number of analyzed cells (N). However, when datasets are large calculation of trajectories starting from every cell can be impractical. In this case, trajectories can be calculated starting from a subset of cells that are selected to be representative of all regions and subsets in the population. We find that 100-1000 trajectories are sufficient for accurate determination of branch points and different developmental paths in all samples we have studied to date (Fig. S1). Phenotypically diverse starting cells, representative of all subsets in the population, are identified by selecting one cell from each of a number of T K-means clusters. The number of clusters is equivalent to user defined number of trajectories (T). Thus, the *trajectory space* matrix (M) can be defined as $M = (d_{ij})$, $i \in [1, N]$, $j \in [1, T]$, and each cell is described by a vector of Wanderlust distances $c_i = (d_1, d_2, d_3, \dots, d_j)$, $j \in [1, T]$. MATLAB and R package implementation of tSpace and accompanying documentation are available online (<https://github.com/hylasD/tSpace> & [https://github.com/hylasD/MATLAB version tSpace](https://github.com/hylasD/MATLAB_version_tSpace)). A tutorial on application of tSpace is available online (<http://denisdermadi.com/tSpace-trajectory-inference-algorithm>) and as a vignette in R package.

Robustness and effect of parameters

To address the robustness of tSpace to parameter selection, we examine (i) the effects of Wanderlust vs Dijkstra trajectory calculation for different T (trajectory number) (Fig. S1); and (ii) the effects of varying K (L is maintained at 0.75K as default in these examples) and distance metric (Fig. S2) on tSpace projections of thymic T cell data. 100 trajectories were sufficient to visualize developmental branching in our thymocyte dataset (Fig. S1A-B, N = ~ 95 000 cells) and scRNAseq (Fig. S1C, N = ~3500 cells); with only minor refinements in cell positioning with 1000 trajectories (Fig. S1).

Furthermore, we compare the effect of the varying K, L and different metrics (Euclidean vs Pearson, Fig. S2). The tSpace output is expected to be influenced by K, the number of neighbors in the KNN graph: Large K, by increasing the number of 'paths' between cells, can lead to unwanted connections (short circuits) to developmentally more distant or altogether unrelated cells. Conversely, if K is too small the neighborhood graph will be unconnected. We find that developmental relationships predicted for thymocytes by tSpace are surprisingly robust over a range of K (from 15 to 100 in Fig. S2). At K = 100, however, the positioning of terminal Tregs is distorted, with some Tregs beginning to form a bridge to early developmental stages (Fig. S2, asterisks & insets). Thus, we suggest a default to K ~ 25. The parameter L defines the subset of K connections around each start cell to be used for sub-graphs; thus, L must be < K. The ratio of L to K determines the independence of sub-graphs that are averaged to further reduce the contribution of short circuits. We suggest L/K ~ 0.75, a value that works well in all datasets we have analyzed.

tSpace implements Euclidean, cosine or Pearson correlation metrics to define local distances between cells. Comparison of tSpace using Euclidean and Pearson metrics is illustrated in Fig. S2.

tSpace running times

Running times of tSpace will be impacted by the size of the dataset and number of calculated trajectories (parameter T) as seen in Supplemental Table 1. Analysis of the *C. elegans* (86,024 cells) failed when we changed parameter T = 5000 trajectories indicating that tSpace have memory limitations, however as demonstrated in Fig. 4 D-G, even if tSpace calculates as few as 50 trajectories cells are ordered in correct developmental time (Fig. 4 D). However increased number of calculated trajectories (already 100 and 500) improves resolution and connection between the cell types (e.g. compare intestine cells, marked with an arrow Fig. 4 D-F).

Supplemental Table 1. Running times of tSpace in minutes for different type of data sets in number of cells, variables and trajectories, related to Figure 1.

| Data set | Cell no. | Variable no. ¹ | Trajectory no. | Waypoint no. | Time [min] | |
|--------------------------------|----------|---------------------------|----------------|--------------|------------|--------|
| | | | | | R | MATLAB |
| B cell | 17 956 | 26 | 100 | 20 | 40 | 19 |
| B cell | 17 956 | 26 | 1000 | 20 | 289 | 198 |
| intestine | 3521 | 2419 | 100 | 20 | 90 | 52 |
| intestine | 3521 | 2419 | 1000 | 20 | 927 | 393 |
| intestine | 3521 | 40 | 100 | 20 | 6 | 2 |
| intestine | 3521 | 40 | 1000 | 20 | 324 | 23 |
| T cell | 94 807 | 12 | 100 | 20 | 313 | 85 |
| T cell | 94 807 | 12 | 1000 | 20 | 1621 | 860 |
| <i>C. elegans</i> ² | 86024 | 75 | 50 | 20 | 147 | NA |
| <i>C. elegans</i> ² | 86024 | 75 | 100 | 20 | 216 | NA |
| <i>C. elegans</i> ² | 86024 | 75 | 500 | 20 | 811 | NA |
| <i>C. elegans</i> ² | 85024 | 75 | 1000 | 20 | 1515 | NA |
| <i>C. elegans</i> ² | 85024 | 75 | 5000 | 20 | Memory out | NA |
| <i>C. elegans</i> ² | 86024 | 50 | 50 | 20 | 115 | NA |
| <i>C. elegans</i> ² | 86024 | 50 | 100 | 20 | 146 | NA |
| <i>C. elegans</i> ² | 86024 | 50 | 500 | 20 | 564 | NA |
| <i>C. elegans</i> ² | 85024 | 50 | 1000 | 20 | 1002 | NA |

¹ variables were measured phenotypic markers (FACS/mass spectrometry) for B and T cell data, variable genes and the first 40 principal components for small intestine scRNAseq

² *C. elegans* datasets was calculated on a computer with 48 cores and 254GB memory

Comparison of tSpace and other trajectory inference algorithms

We compared the output of tSpace with p-Create, Monocle2, diffusion maps, PAGA, and UMAP (Fig. S10-S14), all trajectory inference algorithms designed to be agnostic towards different single cell technologies, and not require any additional input or supervision from the user (e.g. estimates of the relative rates of proliferation and loss, start or end cell points). Each of the algorithms provide different outputs, therefore it is hard to have standardized mathematical comparison, however we use known

biology to determine (i) how faithfully each one of them reconstructs developmental relations, (ii) how deterministic output is and (iii) sensitivity of the algorithm to transient and rare populations.

p-Creode was originally developed for mass cytometry therefore we hoped to find it easy to apply to our FACS and mass cytometry datasets. We found p-Creode, despite the existence of the available tutorial, difficult to optimize due to striking changes if parameters are slightly altered as illustrated in Fig. S10 (T cell) & S11 (B cell). Moreover, in order to run p-Creode, both of our data sets (94,807 T cells, or 17,916 B cells) had to be downsampled below 14 000 cells, impacting detection of rare cell populations (e.g. in Fig. S10 Foxp3+ Tregs are not separated as a branch of CD4 T cells, and in Fig. S13 plasmablasts are never detected). The inconsistent output of cell cluster relations and marker intensities produced by p-Creode, in our opinion, would significantly hinder interpretation of novel datasets. These observed problems are in fact consistent with the original p-Creode publication, where the authors show the algorithm yields quite different outputs when data is re-sampled (Parks et al., 2006).

In contrast, tSpace is highly reproducible and robust. Run on the same sample and with the same parameters, it is nearly deterministic: the only non-deterministic component of the algorithm is the selection of waypoints. It is also robust to the selection of tSpace parameters (Figs. S1 & S2). It is significant that Herring *et al.* independently applied p-Creode to a thymocyte flow cytometry dataset and (just as in our own test) p-Creode failed to define Treg branching. Again, this likely reflects the unavoidable problems associated with the requirement for clustering or downsampling.

Monocle2 failed to order T and B cells in a developmentally meaningful way (Fig. S12A, Fig. S13A). In addition, Monocle2 also required downsampling of both datasets, probably due to memory reasons. Monocle3 provides trajectory inference as one of its applications by utilizing UMAP to embed cells in 2D/3D and finally fitting principal curves using that embedding. In complex looping datasets 2D or even 3D embeddings are likely to position independent branches or pathways on top of each other. Thus, in contrast to Monocle3 we prefer to select pathways/trajectories of cells manually, using at least the first 3-4 tPCs. Due to Monocle3 reliance on UMAP please see further below UMAP comparison with tSpace.

Diffusion Maps ordered cells comparable to tSpace providing developmentally correct relations (Fig. S12B, Fig. S13B). However, diffusion maps do not reveal multiple branches as clearly as tSpace (please compare Fig. 1B and Fig. S12B). Also, in the diffusion map projection IgM memory B cells are positioned as an intermediate state from naïve B cells to memory IgA and IgG B cells, instead of stemming as a separate branch from naïve B cells (the latter expected from known biology). The limited ability to reveal multiple branches is consistent with the original report (Butler et al., 2018), in which the authors claim that for multiple subsequent branches, diffusion maps should be applied iteratively. Interestingly, the first diffusion component, for both data sets (mouse T and human B cell), did not reflect developmental relations (Fig. S12C & Fig. S13C).

Diffusion maps (DM) is faster algorithm than tSpace, however empirically we show (Fig. S12 B-C, Fig. S13 B-C, Fig S14 D-E), what the authors of DM and diffusion pseudotime (DPT) state in the article, that DM & DPT perform well for one bifurcation but for more complex structures identification of branches has to happen iteratively, making it impractical for complex data such as mouse intestine or *C. elegans*. A great example is a complex intestinal dataset, which contains ample number of enteroendocrine cell subtypes, as seen in tSpace representation (Fig. 3A). DM does not even resolve fully enterocytes, nor shows any of the enteroendocrine cell types (Fig. S14 D-E). Here we would like to emphasize that DM algorithm lends its time efficiency to matrix operations, which may have its limitations in large datasets. Indeed, as implemented in R, diffusion maps cannot handle datasets large as *C. elegans*. Function requires a distance matrix created from the gene expression data, which in this case would require a matrix with dimensions of 86024x86024; unlikely to be handled memory-wise by a cluster of 48 cores with 254 GB memory, even less so by a personal computer.

Here we would like to clarify that DM-DPT and Wishbone, yet another algorithm related to tSpace because of use of Wanderlust, both use diffusion maps embedding. Namely Wishbone applies Wanderlust on kNN graph, constructed using diffusion maps embedding, while tSpace applies Wanderlust on subgraphs (keep L out of K nodes, each subgraph can be seen as a subset of “random

walks”) of a kNN graph constructed using all informative principal components of gene expression space. Haghverdi *et al.* claim that use of Wanderlust provides unreliable geodesic distances for complicated manifolds containing turns, multiple branching and larger amount of noise (Haghverdi et al., 2016). However, we demonstrate using real world complex data, such as *C. elegans*, that tSpace, estimates cell positions remarkably well and concordant to the real cell developmental time (Fig. 4 B, D-F), separates all major lineages in *C. elegans* and retains precision in the early developmental stages (Fig. S9).

Partition-based graph abstraction (PAGA) relies on existing algorithms to unveil data manifolds, such as t-SNE or UMAP, (Fig. S12D & Fig. S13D) and the Louvain algorithm to determine clusters (Fig. S12D-F & Fig. S13E-F). The “novel step” in PAGA analyses determines meaningful connections between the clusters and partitions the graph. PAGA analysis of T and B cells is summarized in Fig. S12D-H and S13D-G, respectively. PAGA on T cell data used 13 and 26 clusters. While automatic clustering for the most part agrees with manual gating (Fig. S12E-F & Fig. S13F), it fails to subset DP TCRβ+ CD3ε+ (DP*), SP CD8 recent emigrants in T cell dataset, and is not always concordant with expert manual gating, which remains the gold standard in immunology. PAGA analysis of T cells (Fig. S12G), even after optimization, does not connect DN2 and DN3 (cluster 10) to the rest of the T cells, and connects SP CD4 (cluster 7) and SP CD8 (cluster 9) via Treg (cluster 11). Treg should branch from CD4 cells. [Interestingly, at very high K of 100, tSpace begins to make this aberrant connection via Treg as well (Fig. S2; this is discussed as an issue with tSpace parameter selection earlier).

Increasing the number of clusters in PAGA T cell analysis from 13 to 26 only fractionates DP T cells and does not improve connectivity between cell populations (Fig. S12H). PAGA analysis of B cells used 23 clusters and (Fig. S13G) connects clusters similarly to diffusion maps: blood naïve B cells start in cluster 4 and differentiate via clusters 12, 3, 0, 6, 1 into memory IgM cells (cluster 13), or IgG memory (cluster 7), and IgA memory (cluster 9). Memory IgM cells are not connected to germinal center (clusters 5, 14 and 2). Cluster 15, mix of IgA, IgG memory cells connects to cluster 2, which is a mix of IgA GCB, and to lesser extent centroblasts and GCB IgG B cells, already suggesting discrepancy to known biology in which GCB IgA or IgG cells should be differentiated out of centroblasts (clusters 5 & 14). Overall, PAGA does not agree with known biology completely, which can be problematic. Additionally, as much as the abstraction step in PAGA can be useful as a summary, it can also reduce a biologist’s ability to explore the data and isolate fine structures or features seen only in single cell/single point representations. This is amply illustrated in the confusing connectivity of abstracted relationships by PAGA in our tonsil dataset. In contrast, the tSpace projection allows visualization of predominant paths and branches but retains positioning of individual cell intermediates that rather intuitively illustrate alternative paths suggested by the data.

Uniform Manifold Approximation and Projection (UMAP) of T cell (thymus) and B cell (tonsil) data unveils the manifold (Fig. S12D & Fig. S13D) and aligns cells in developmentally meaningful ways, similarly to tSpace. However, tSpace analysis of intestinal single cell RNAseq outperforms UMAP in preservation of cellular developmental relations (UMAP clusters cells appropriately but does not retain the lineage relationships. Fig S14A & B).

Although UMAP applied conventionally to phenotypic profiles does not perform as well as tSpace, UMAP offers an interesting tool for visualization of tSpace relationships. Please see and compare UMAP embedding of cells based on tSpace trajectories (Fig. S6 and S14C) and contrast it with UMAP embedding of phenotypic profiles (Fig. S14A & B). UMAP does a good job of reducing trajectory space to two dimensions while preserving relationships and branching.

We also analyzed the intestinal scRNAseq data with diffusion maps, Monocle2 and UMAP (Fig. S14). Overall, the diffusion map algorithm performs well in determining major branches but fails to unveil early differentiation steps or sub-branching within the absorptive and secretory branches. Additionally, rare populations e.g. sLEEP cells (express *Neurog3*), are not easily detectable (Fig. S14D-E). UMAP, depending on the minimum distance parameter (minimum value ~0, maximum value 1) will either show separated clusters (Fig. S14A, min_dis = 0.1) or gradual connections between them (Fig. S14B, min_dis = 1). Contrary to diffusion maps, UMAP separates rare sLEEP cells from goblet cells (Fig. S14B, arrow). However, even with the maximum value for minimum distance parameter, UMAP embedding of cell

profiles is not able to show the known connection between sLEEP cells and rest of the enteroendocrine subsets.

Conclusions

Taken altogether, tSpace visually represents all intestinal developmental stages more equally than diffusion maps and detects multiple branches better than diffusion maps or UMAP alone. Existing algorithms by virtue of clustering (or use of force directed graphing) also limit the ability to visualize the existence of individual or scattered cells between major paths, which may reflect the diversity of differentiation sequences possible within a population. We believe a major reason for the strong performance of tSpace is the Wanderlust algorithm for calculation of distances, which distinguishes it from all the other methods here. The representation of cells in the dense matrix of trajectory space rather than the sparse matrix of KNN graph space also adds to the robustness of the method.

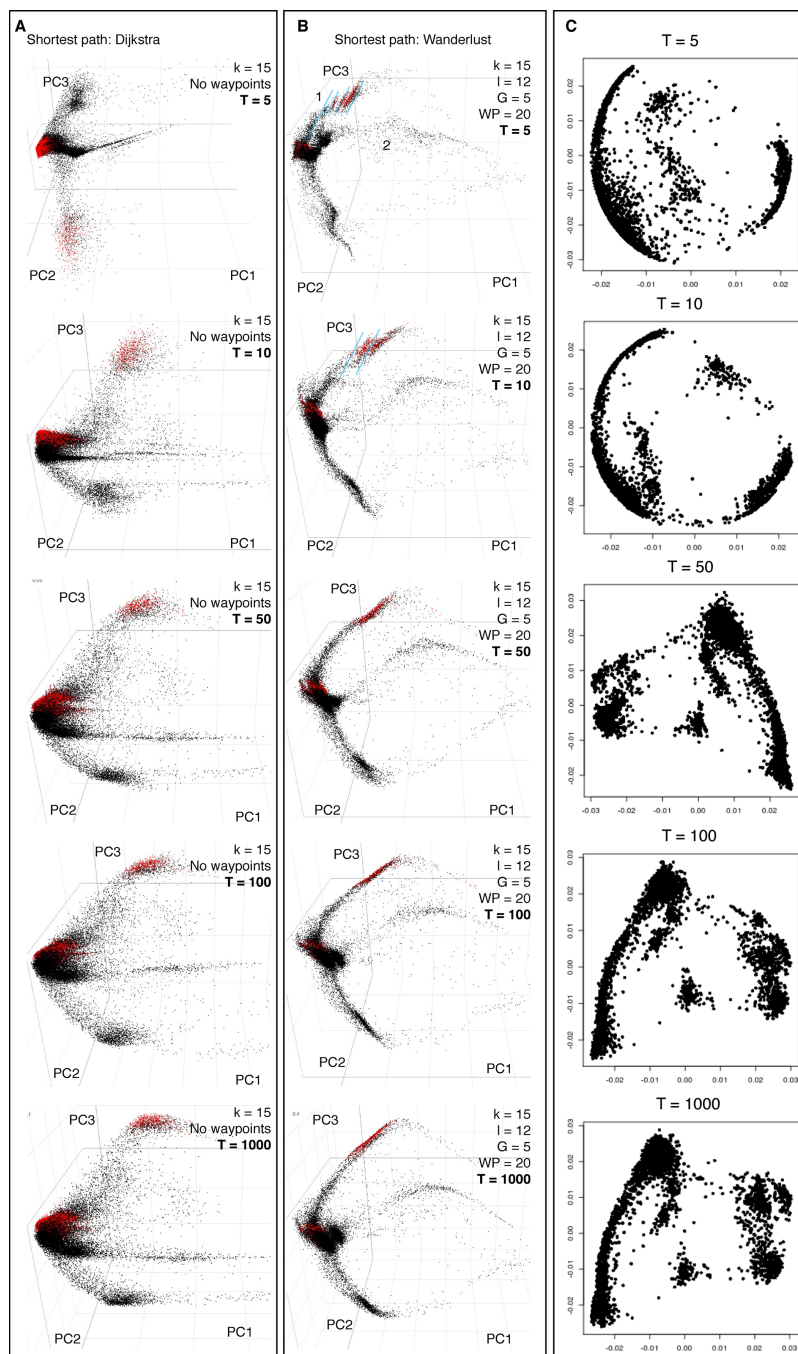


Fig. S1. Effect of Wanderlust and of number of calculated trajectories (T) on tSpace output, related to Figure 1: A-B thymic T cell data and c scRNAseq of mouse intestine. A tSpace analysis using the Dijkstra algorithm to define distances (Dijkstra, 1959). The Dijkstra algorithm calculates shortest distances between cells within a KNN, but without the subgraph averaging or waypoint optimization implemented in Wanderlust. An increase in calculated trajectories slightly improves the shape of data. Mathematically, tSpace with Dijkstra is similar to the isomap (Tenenbaum et al., 2000). **B** tSpace analysis of thymocyte FACS data using Wanderlust (Bendall et al., 2014). Use of Wanderlust to define trajectories results in tighter and more distinct branches (compare right to left panels). Increasing the number of calculated trajectories improves the definition of developmental sequences as well; but the position of cells in trajectory space (in the tSpace output) tends to stabilize between 100 and 1000 trajectories in experimental datasets. Breaks between cells (blue lines) seen with very low T (5 or 10 calculated

trajectories) reflect the influence of waypoints. For orientation, we labeled in red DN3 and DP T-cells. **C** tSpace on intestinal scRNAseq dataset illustrating stabilization of developmental relationships between 100 and 1000 T. For visualization of cell relationships in trajectory space, PCA embedding is used here.

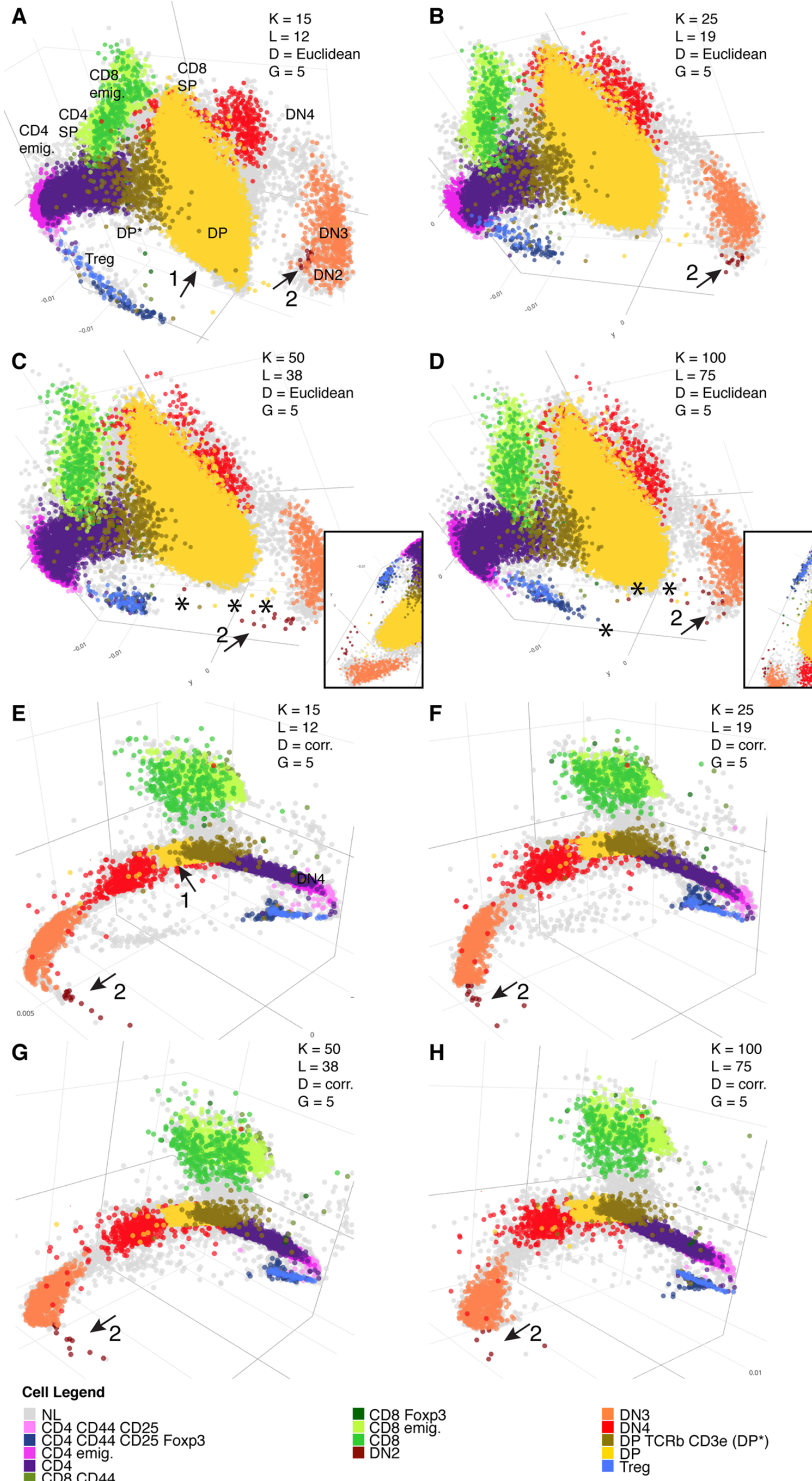


Fig. S2. tSpace output as a function of K , L , and distance metric (Euclidean vs. Pearson correlation), related to Figure 1 A-D PCA embedding of trajectory space matrices with different parameters using Euclidean distance. E-F PCA embedding of trajectory space matrices with different parameters using Pearson correlation. Pearson correlation tightens populations and branches compared to Euclidean (arrow 1). Overall results are robust to a range of parameters, but smaller K reveals more local details and structures of the manifold, and prevents dispersion of cells (e.g. see small DN2 subset, arrow 2). With high K (Euclidean metric), some DN2 cells begin to form a bridge inappropriately towards

terminally differentiated Tregs (asterisks and inset panels showing different angle). The choice of distance metric (Euclidean vs Pearson correlation) affects the shape of the manifold but developmental relationships are retained. The numbers of trajectories ($T = 250$) and subgraphs ($G = 5$) were constant. Results with 15 subgraphs were similar (not shown). NL – not labeled, DN – double negative, DP – double positive.

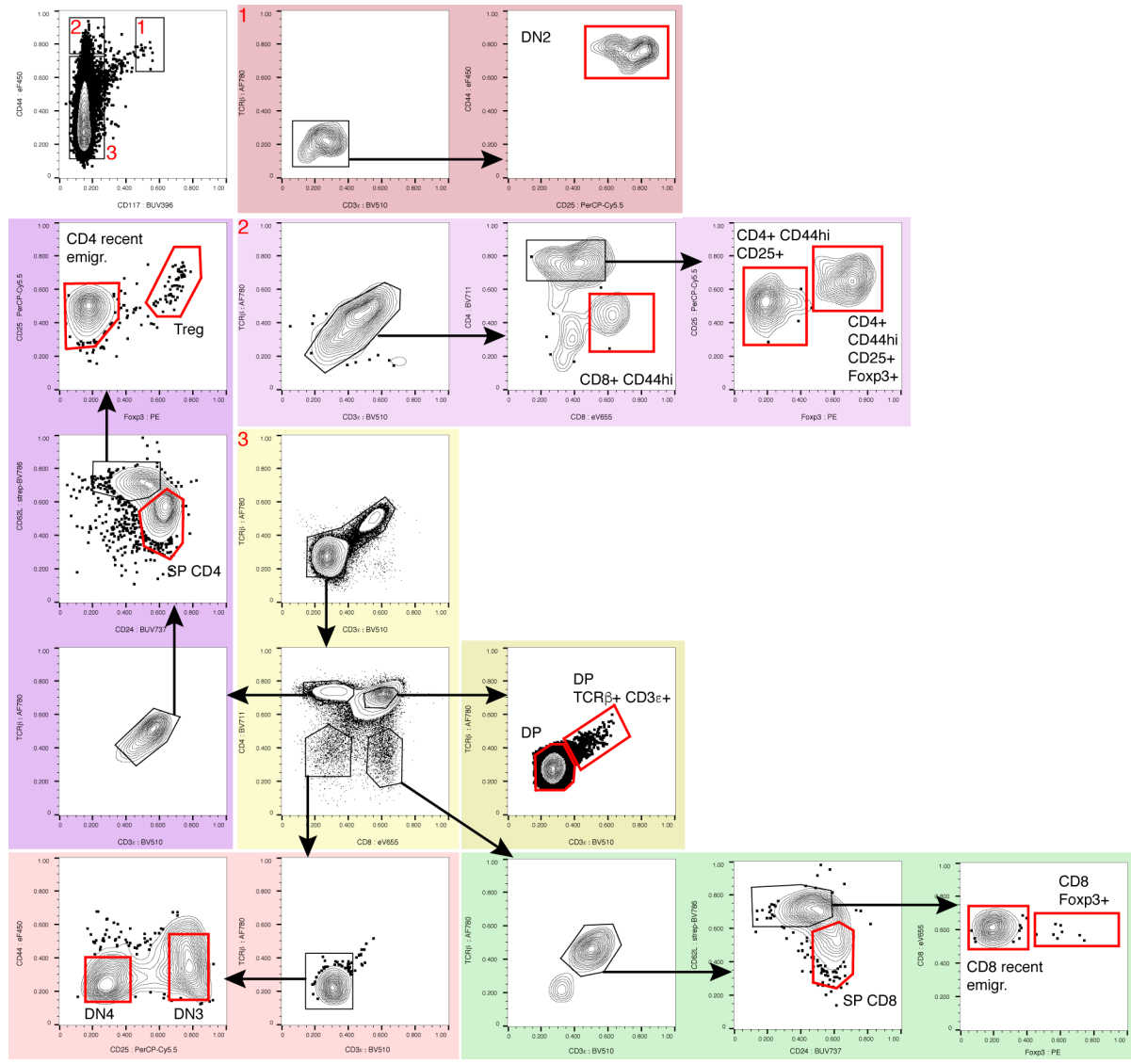
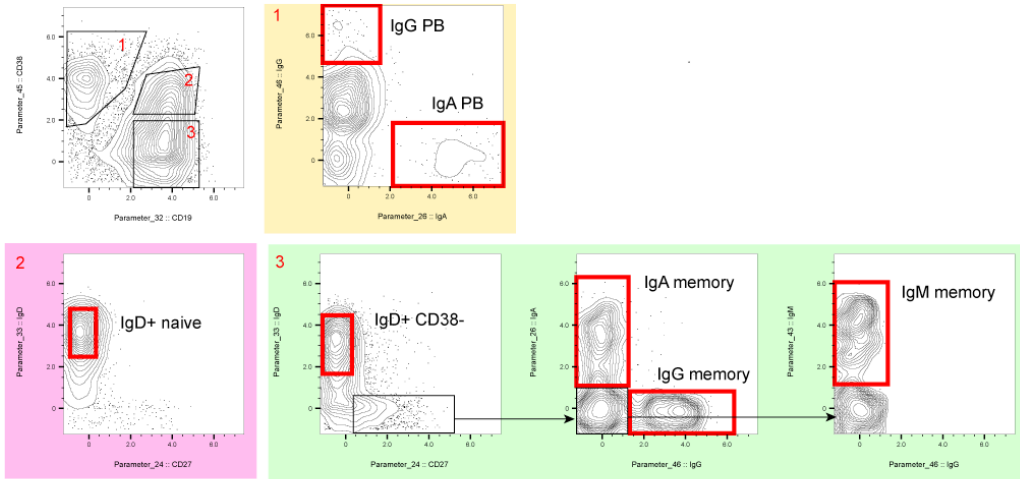


Fig. S3. Manual gating strategy for conventional definition of T cell subsets in the thymus, related to Figure 1. Red gates mark final cell populations used for highlighting in tSpace visualization. Arrows mark flow of the gating.

Blood



Tonsil

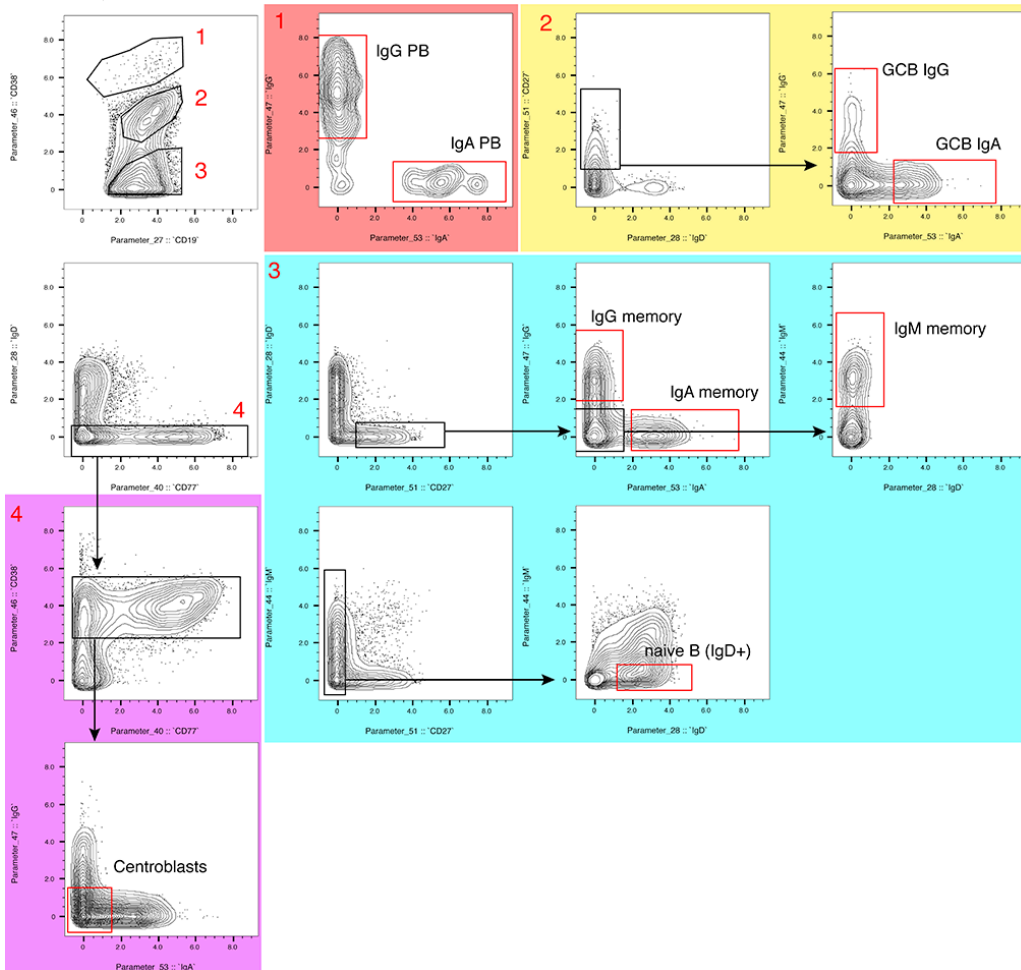


Fig. S4. Manual gating strategy for tonsil B cell subsets, related to Figure 2. Red gates mark final cell populations labeled for visualization in tSpace analyses. Arrows mark flow of the gating.

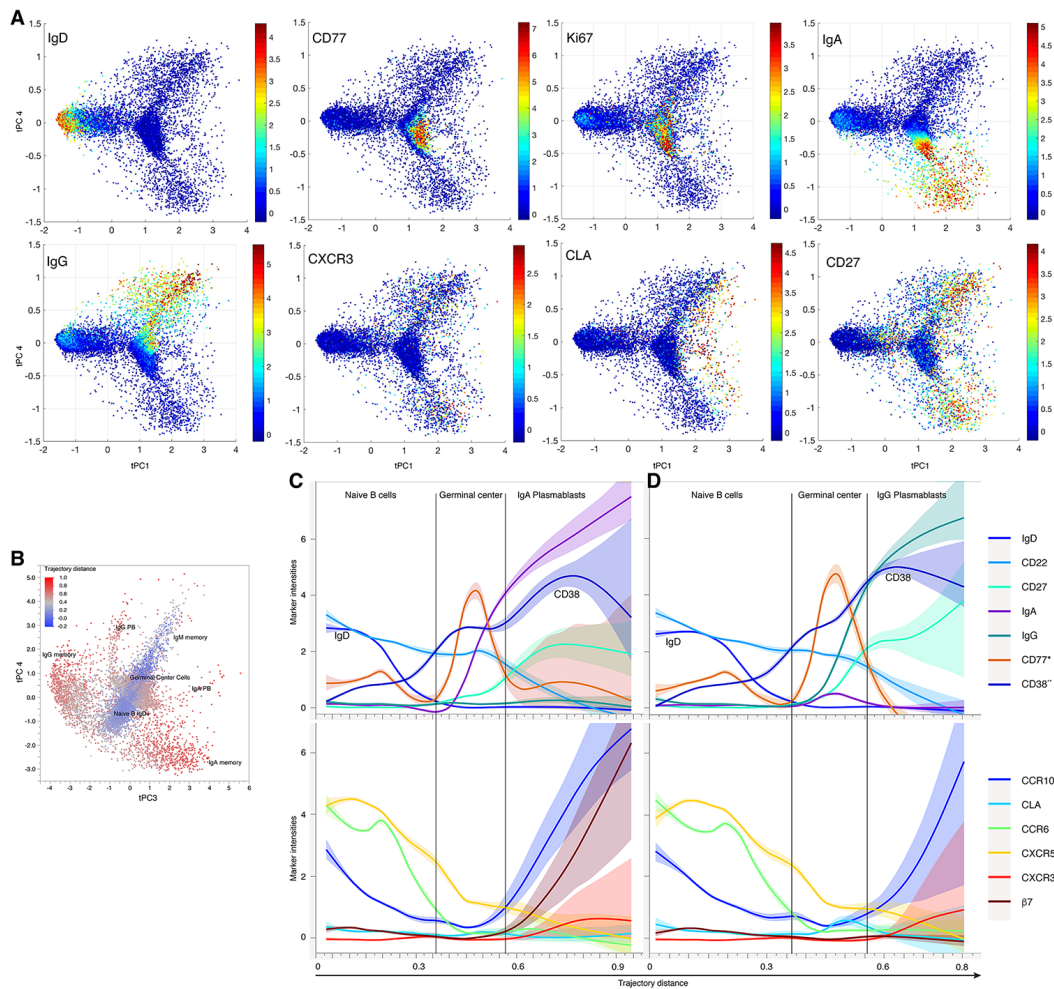


Fig. S5. tSpace reveals tonsil B cell development, related to Figure 2. A Principal component embedding of cells in trajectory space: principal components tPC1 (delineating naïve to GC and mature memory transition) vs tPC4 (delineating pathways to IgG vs IgA expression) are plotted. Coloring based on cell staining for the indicated markers. **B** tSpace projection using tPC3 and tPC4 with cells colored by distance from naïve B cells within trajectory space. **C-D** Changes in phenotypic markers and trafficking markers along B cell maturation trajectory from naïve B cells to IgA (C) or IgG (D) plasmablasts.

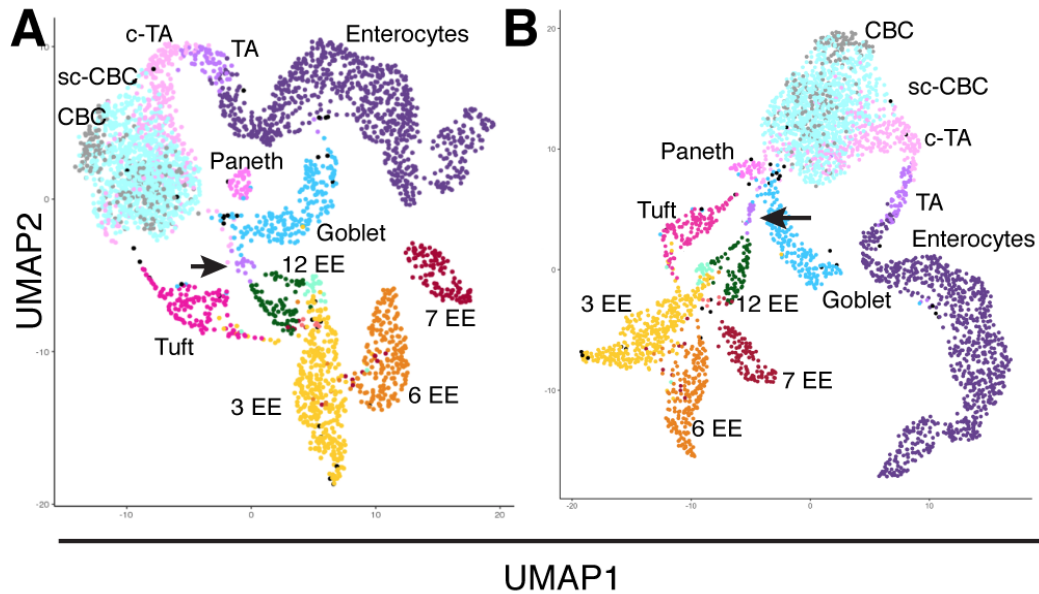


Fig. S6. UMAP embedding of tSpace distance matrix, related to Figure 3. Most developmental branching relationships are retained in 2D UMAP embedding of the dense tSpace distance matrix (**A-B**). UMAP of tSpace matrix with $\text{min_dis} = 1$ **A** Euclidean and **B** Manhattan distance. siEEP cells are marked with an arrow. TA – transit amplifying, CBC – crypt base columnar, sc-CBC slow-cycling CBC, EE – enteroendocrine cells.

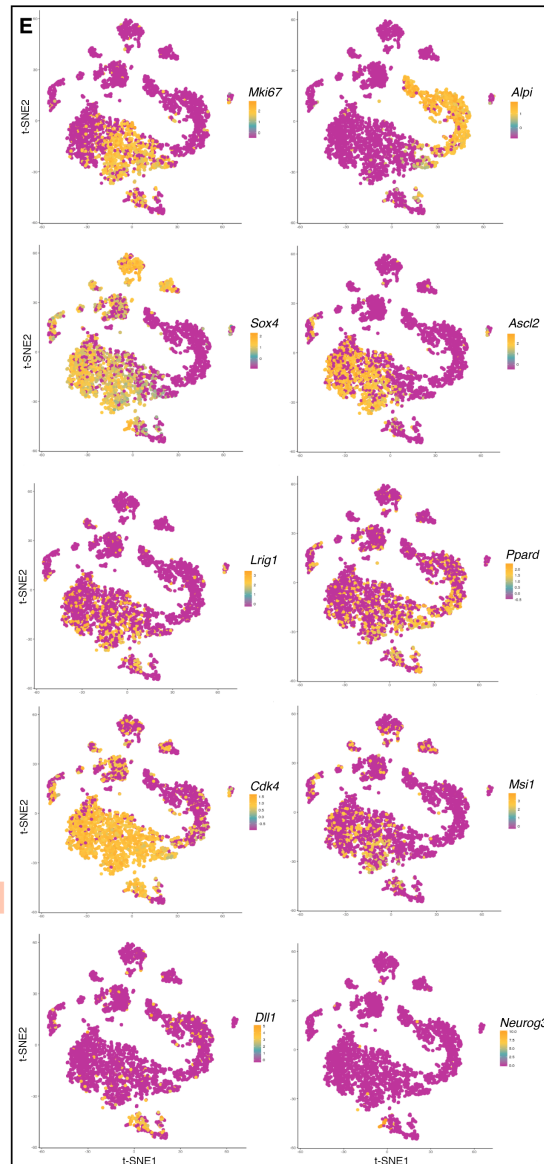
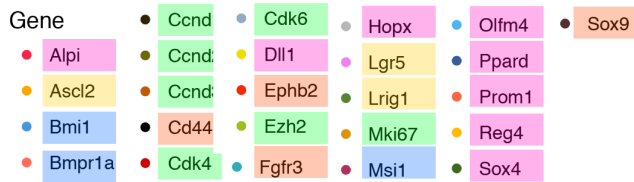
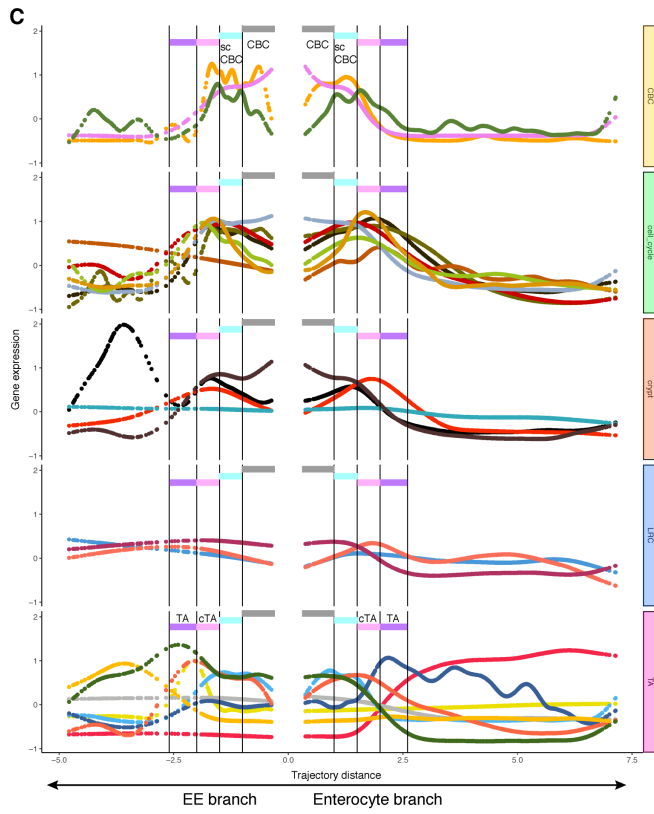
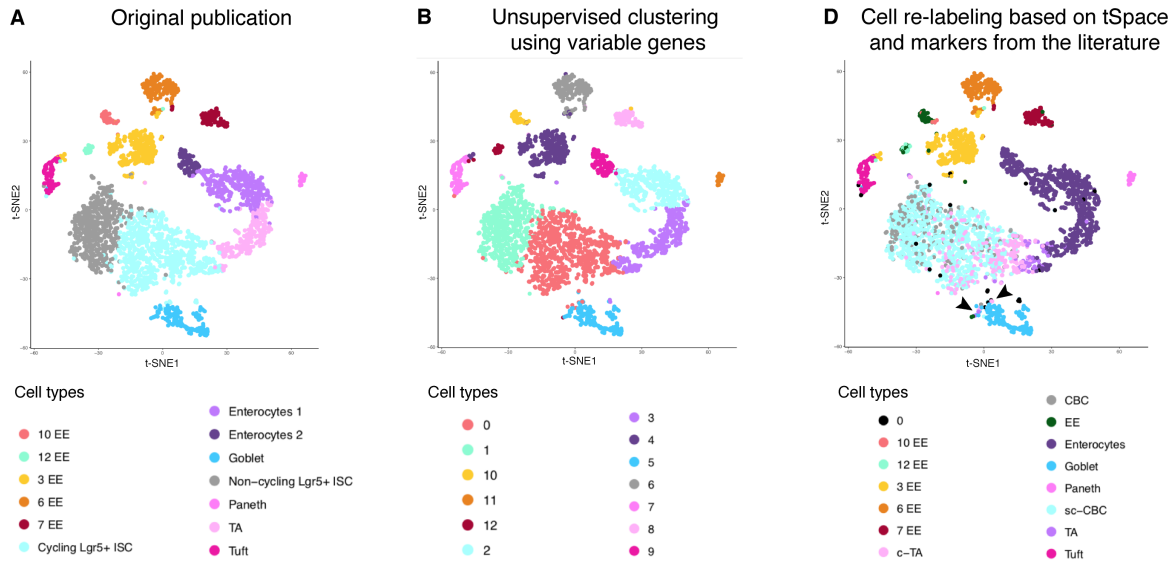


Fig. S7. Proposed labels for the cell populations within the intestinal crypt based on developmental distance (distances along isolated cell paths from tSpace) and subset markers, related to Figure 3. A t-SNE visualization of cell populations, with the original labels². **B** t-SNE visualization of cell populations determined using all variable genes and unsupervised clustering provided in Seurat package. Annotation of subsets is from Yan *et al.* based on unsupervised clustering and association with known intestinal populations. **C** Smoothed expression of markers⁵ that define the crypt zone (orange), cell cycle and proliferation (green), and populations (CBC, TA) associated with the intestinal crypt along isolated trajectories: these markers were used to re-evaluate cell designations, identify putative label retaining cells (LRC, blue), crypt base columnar (CBC, yellow) and transit amplifying (TA, pink) cells. We used peaks of expression of multiple subset-associated genes to define specific cell types. Vertical black lines are suggested boundaries between cell populations. The combination of proliferation markers (green) with CBC and TA cell markers allows separation of slow cycling CBC and cycling TA cells. *Dll1* and *Sox4* are exclusively expressed by TA within the EE branch, while *Alpi* gene and *Ppard* are specific to TA within the enterocyte branch. The cell identities defined provide landmarks for orientation, but tSpace visualization emphasizes that these populations actually exist in the tissue as part of a developmental continuum. **D** t-SNE visualization of the newly tSpace-defined cell populations within the crypt. Arrows point to the SLEEP cells, delineated in trajectory space and known to exist in intestinal crypt, but which are not defined by conventional t-SNE, SPADE or unsupervised clustering (Seurat) but rather clustered with goblet cells. Cells labeled with 0 are all cells that did not fall into our vertical gates. **E** Selected markers from **C** panel, shown in a t-SNE map.

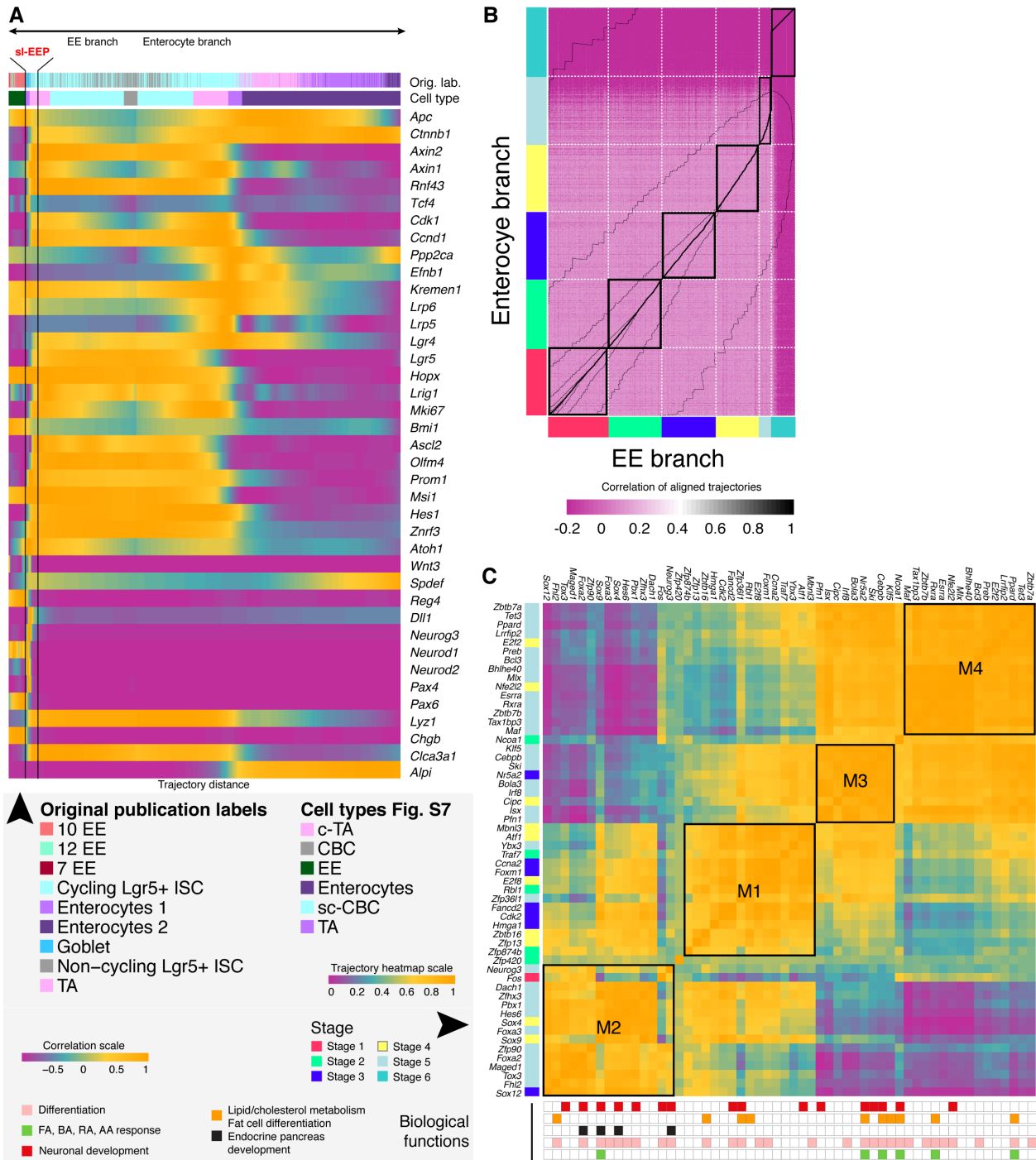


Fig. S8. Trajectory analysis allows identification of candidate TF's and TF modules for enterocyte and enteroendocrine (EE) differentiation, related to Figure 3. **A** Expression patterns of genes that regulate intestinal crypt development reviewed by Hans Clevers, along the isolated trajectories confirm known biology. Cell labels from original publication, our proposed cell labels and short lived enteroendocrine progenitors (siEEP) are marked along the trajectories. Cell stage (Transparent Methods) and cell identities defined here (cell subsets) or in Yan *et al.* (Original labels) are indicated above the heatmap. **B** Heatmap of single cell correlations between two trajectories and alignment cost (thin black contour bars), which is minimal along the thicker black line. Trajectories were aligned with dynamic time warping and subsequently sectioned into 6 segments for further pair-wise comparison (e.g. Stage 1 EE

branch vs Stage 1 enterocyte branch). **C** Co-expression analysis of transcription factors identifies 4 modules (M1-M4) specific for intestinal crypt and cell commitment to enterocyte or EE lineages. Stages shown next to the TFs are the earliest stage when TF was significantly changed in one of the compared trajectories. Biological functions of TFs are highlighted in the panel below the correlation heatmap.

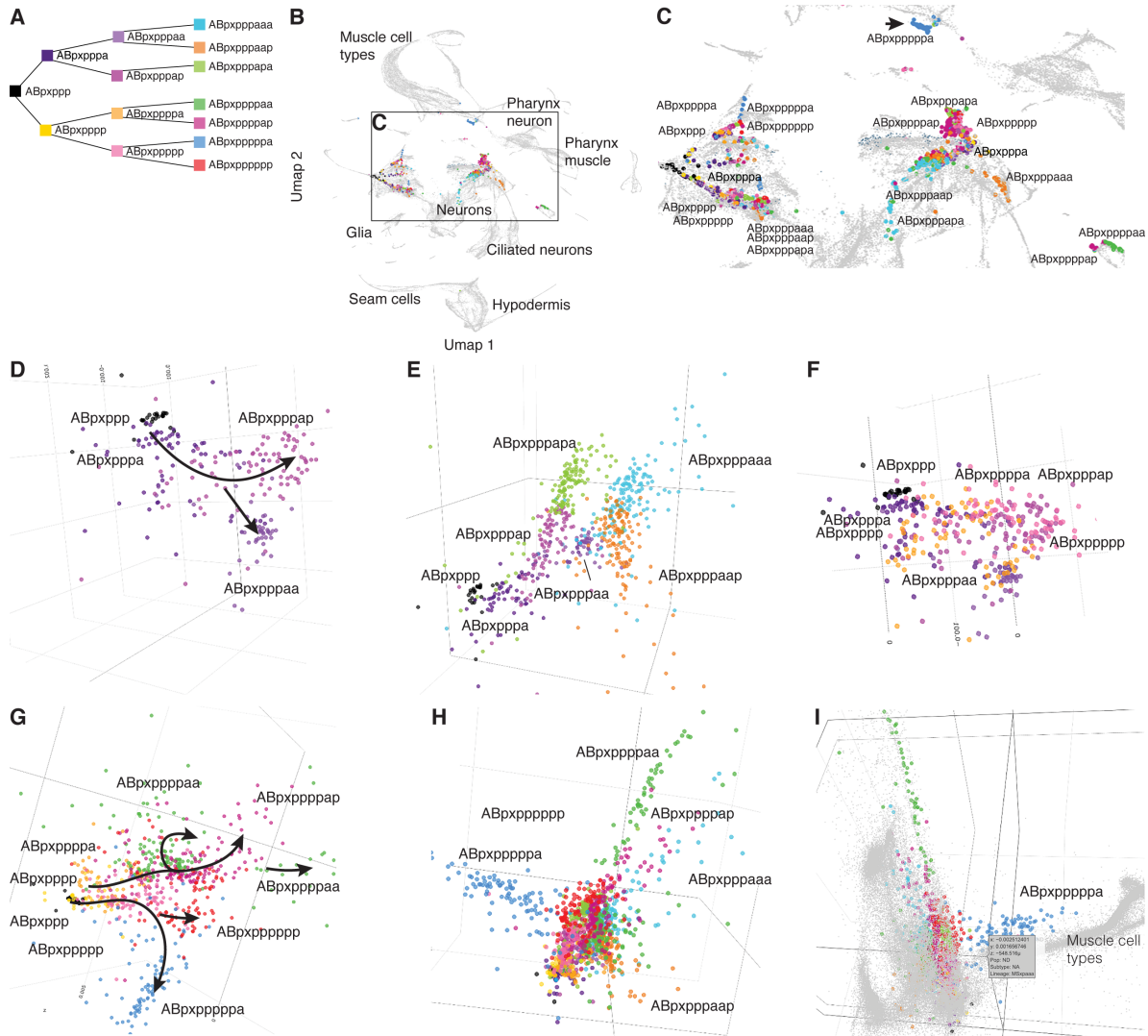


Fig. S9. Benchmarking of tSpace performance using full *C. elegans* dataset demonstrates perseverance of developmental relations even in the early developmental phases, related to Figure 4. **A** Lineage tree of the ABpxppp sublineage descendants [recreated from Fig. 2F in (Packer *et al.* 2019)] a – anterior, p – posterior, x – left or right, as part of the standard nomenclature in *C. elegans* development. **B** and inset **C** UMAP of 75 PCs of gene expression matrix, recreating the UMAP from Packer *et al.*, showing the ABpxppp sublineage scattered and disconnected. In order to create correctly connected sublineage cells Packer *et al.* had to calculate separate UMAP of 8083 neuron, glia, and rectal progenitor cells with embryo time ≤ 250 min unlike **D-I** tSpace calculation using full *C. elegans* dataset. **D-E** The ABpxpppa sublineage cells seen in tPC1, tPC4 and tPC5 are well connected and correctly ordered as expected (**A**). **E** Three cell subpopulations are branched out from the correct cell precursors. **F** tSpace showing ABpxpppa, ABpxpppp and their imminent cell daughters, correctly ordered in developmental time, but visually not separated. **G** The ABpxppppp sublineage cells seen in tPC1, tPC4 and tPC5 are well connected and correctly ordered. **H** tSpace showing full ABpxppp sublineage in tPC1, tPC4 and tPC5. **I** tSpace showing full ABpxppp sublineage in tPC1, tPC4 and tPC5 and terminally differentiated muscle cells. ABpxpppppa (blue) is close to the muscle branch but very distinct unlike in UMAP (see arrow in inset **C**). All muscle cells develop from MS lineage not AB, a fact hard to conclude in UMAP analysis (inset **C**).

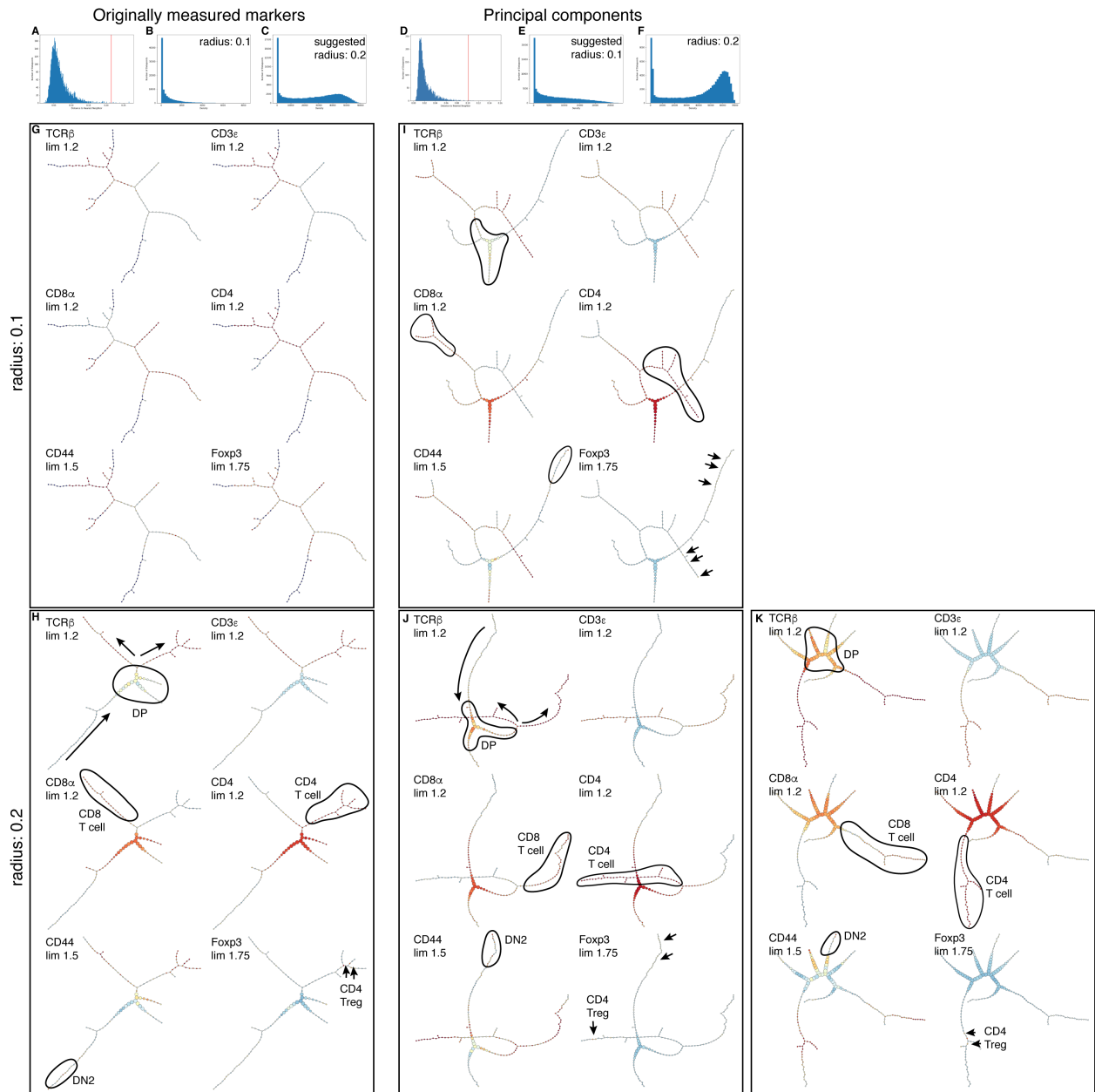


Fig. S10. p-Creode analysis of thymocyte development, related to Figure 1. p-Creode analysis requires downsampling, is inconsistent and sensitive to choice of parameters, related to Figure 1. We ran p-Creode using (A-C, G-H) originally measured markers vs. (D-F, I-K) principal components, which is suggested by p-Creode tutorial. We compared the effect of different values of radius parameter. A and D suggested radius, B, C, E and F cell density graphs for different radius. G, H, I, J different p-Creode analysis depended on selection of radius (G & I radius 0.1, H & J/K radius 0.2) and measured markers vs 5 principal components. K p-Creode analysis of the same T cell data with the same parameters as in J, showing effect of random downsampling on the final output of p-Creode.

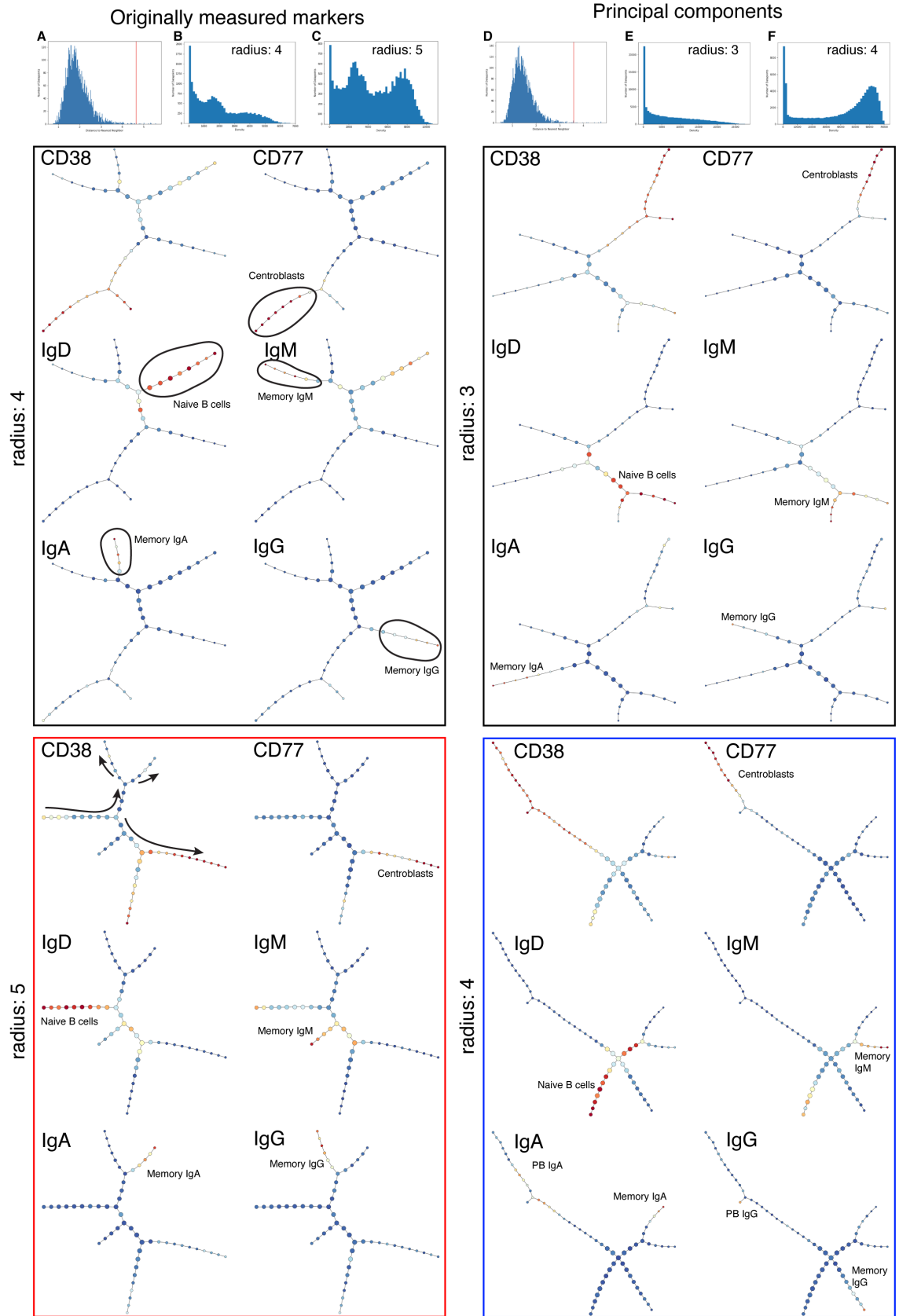


Fig. S11. p-creode analysis of tonsil B cells, related to Figure 2. As for thymocytes, the output is

sensitive to choice of parameters. It also fails to differentiate small populations (plasmablasts), related to Figure 2. p-Creode of tonsil B cells. We ran p-Creode using originally measured markers (left panels) or principal components (right panels), which is suggested by the p-Creode tutorial; and we compared the effect of different values of the critical radius parameter. Red and blue squares mark analyses that are closest to real biology. A and D suggested radius, B, C, E and F cell density graphs for different radius. Four squares show different p-Creode analysis depended on selection of the radius and on use of measured parameters or 15 principal components. Red and blue squares label closest p-Creode solution to the real biology.

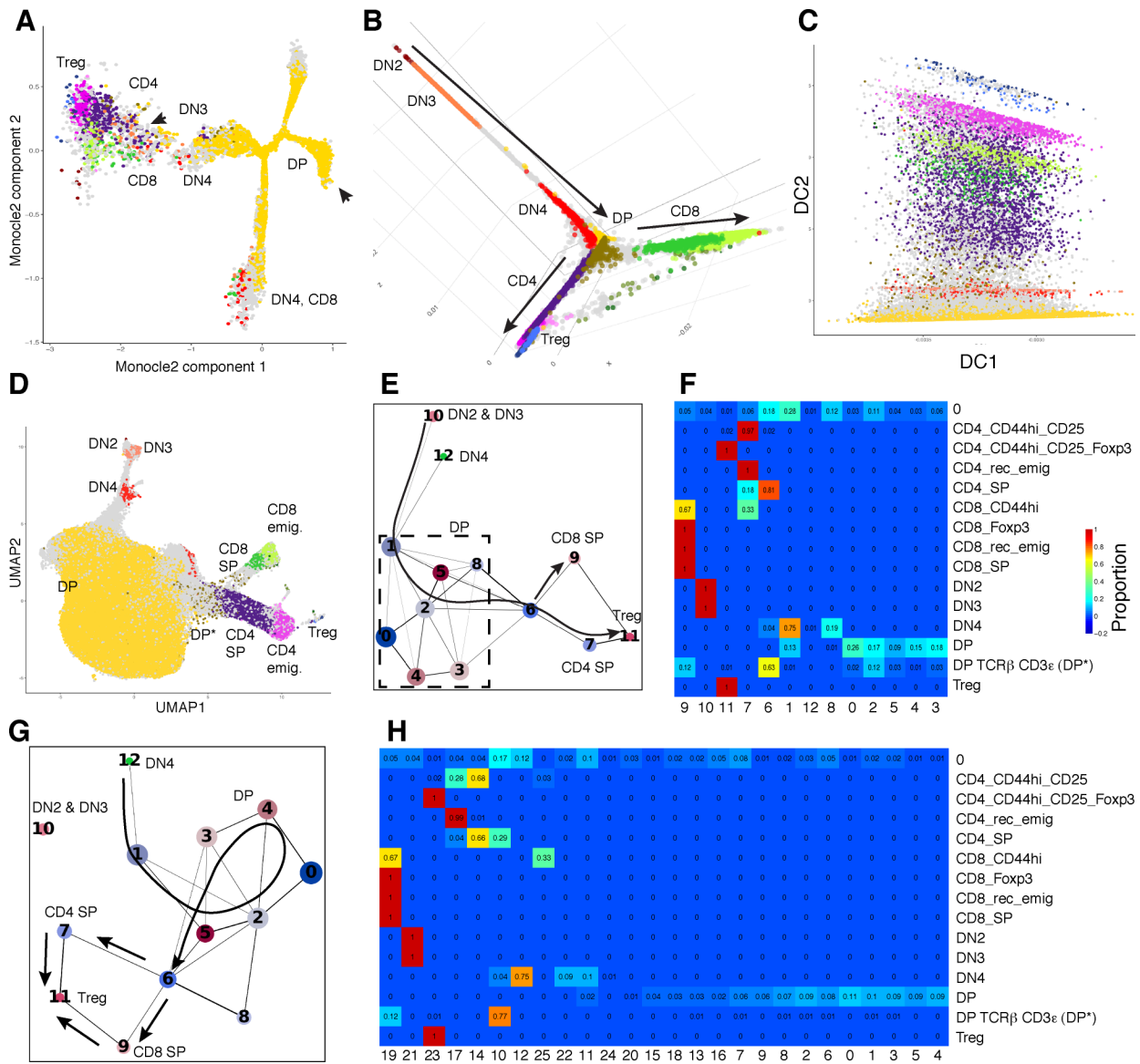


Fig. S12. T cell data analyzed using **A** Monocle2, **B-C** diffusion maps, **D** UMAP and **E-H** PAGA, related to Figure 1. Monocle2 failed to reveal the relatively simple T cell maturation branching. Diffusion maps, if the first diffusion component is ignored (DC1 is shown in panel C) reveals CD4 and CD8 branching as appropriate, however subtle sub-branches (i.e. of Tregs vs conventional CD4 mature T cells) are not prominent. PAGA fails to connect DN2 and DN3 precursors to DN4 and the rest of more mature T cells.

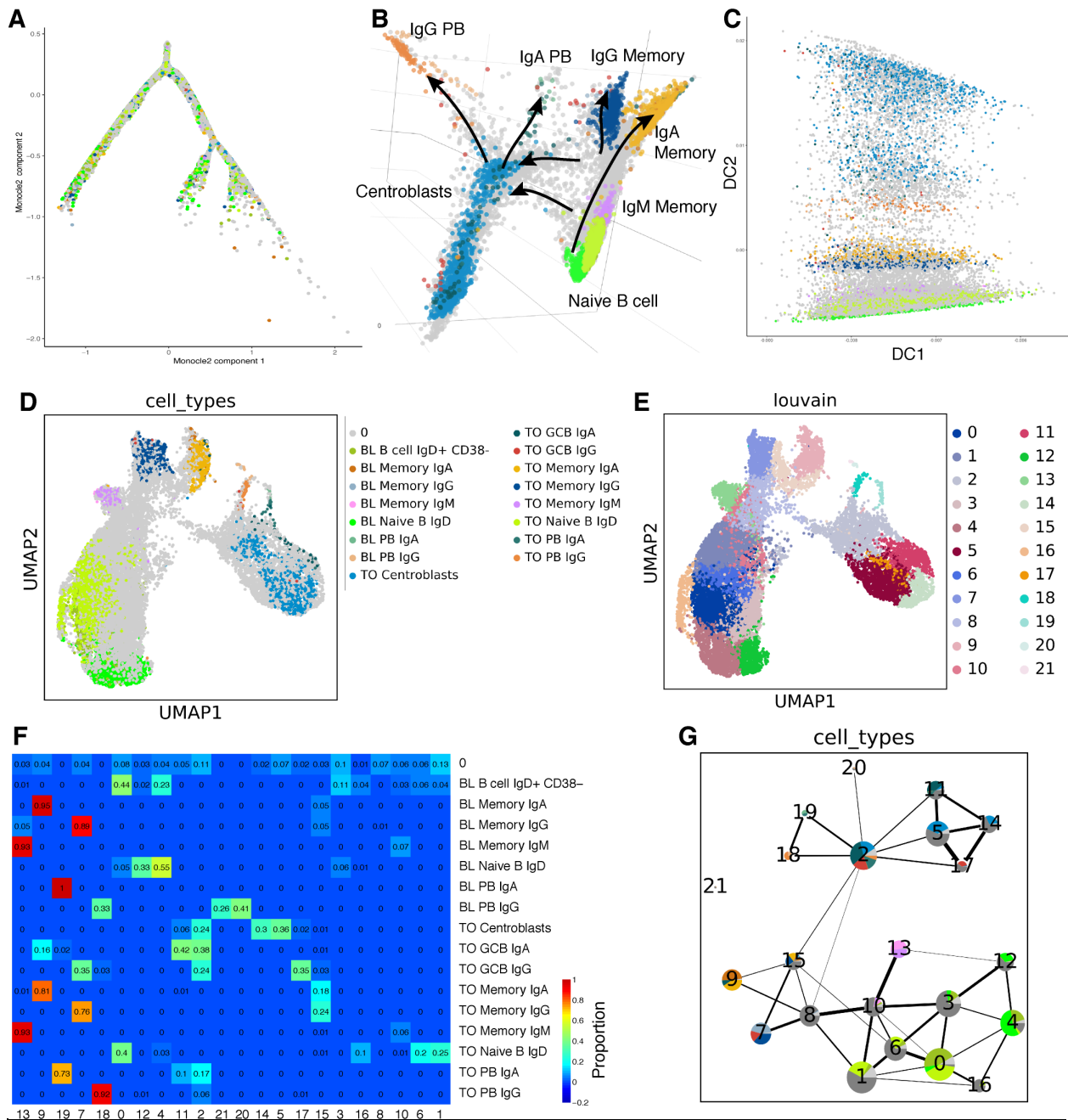


Fig. S13. B cell data analyzed using A Monocle2, B-C diffusion maps, D UMAP, E-G PAGA, related to Figure 2. Monocle2 (A) fails to reveal B cell maturation in tonsil. Diffusion maps (B and C), as in T cell data if the first diffusion component is ignored, reveals central developmental paths in B cell development but positions IgM memory cells along the core path from naïve cells to germinal center cells (GCC); but obscures links from GC to memory cells that are revealed by intermediates in tSpace. UMAP (D) clusters cells obscuring intermediate populations and potential alternative differentiation sequences. PAGA (E-G) connects naïve cells with germinal center (GC) B cells through IgG and IgA memory B cells, without direct connection between naïve and GC.

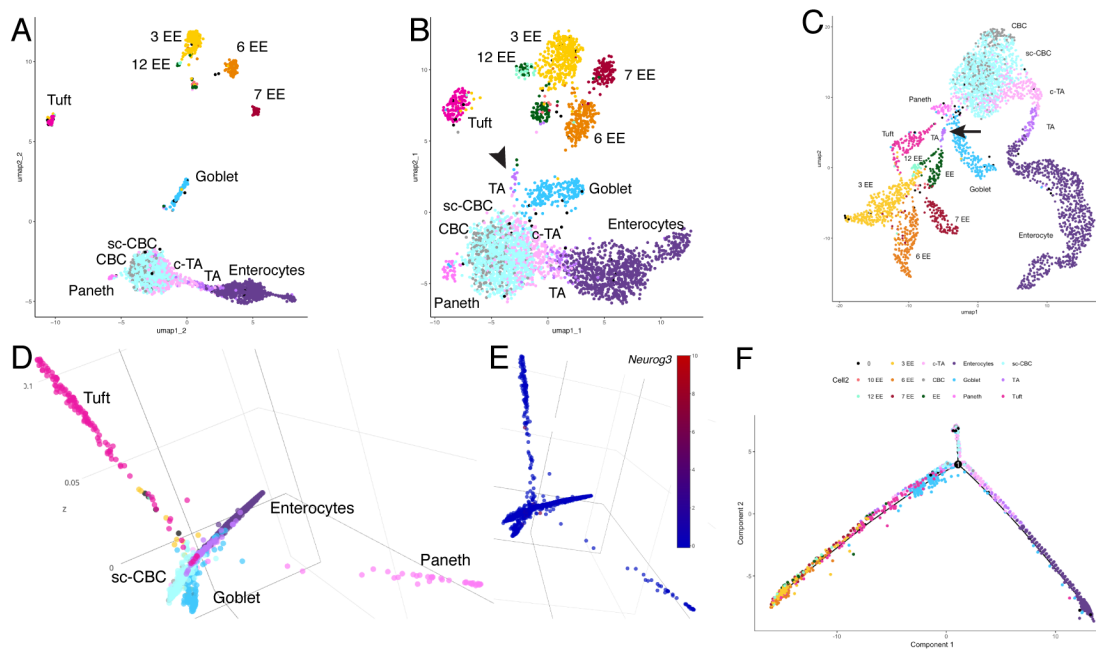


Fig. S14. Diffusion map and UMAP analysis of mouse small intestine, related to Figure 3. A-B UMAP of gene expression with Euclidean distance and different `min_dis` parameters **A** 0.1, **B** 1. UMAP on variable genes, even with the “loosest” setting for `min_dis` (B) fails to connect sLEEP cells to the rest of secretory subsets. **C** UMAP embedding of cell trajectory profiles (tSpace matrix), in contrast to UMAP of gene expression profiles (B), distinguishes sLEEP cells and correctly positions them on a path from stem cells to EE populations, which are well separated in the trajectory space UMAP projection. UMAP embedding, for visualization, of the trajectory space matrix (here and in Fig. S6) and PCA embedding of tSpace (as shown in text Fig. 3A) outperform all of the methods we evaluated. Arrow marks position of sLEEP cells. **D** Diffusion maps with all cell types. Diffusion map branches Paneth, goblet, tuft and enterocytes, while the rest of secretory subsets are not represented, and it fails to show connection between crypt base columnar cells and the rest of differentiated cells, suggesting that diffusion maps perform well in determination of differentiated cells, however **E** *Neurog3* gene expressing cells, described in the manuscript as sLEEP cells are lost in the cloud of early stem cells. **F** Monocle2 reveals only two major branches absorptive and secretory, without any subtleties, and sub-branches of the secretory branch are missing.

Transparent Methods

Lead contact and materials availability

Datasets generated in this study (FACS and CyTOF) are available upon reasonable request. Further information and requests will be fulfilled by the Lead Contact, Denis Dermadi (ddermadi@stanford.edu, denisdermadi.com).

Experimental model and subject details

Animals

Two C57BL/6J male mice, used for thymic T cell isolation and immunofluorescent microscopy, were bred and maintained in the animal facilities of the Veterans Affairs Palo Alto Health Care System, accredited by the Association for Assessment and Accreditation of Laboratory Animal Care. All animal work was approved by the Institutional Animal Care and Use Committee at the Veterans Affairs Palo Alto Health Care System. Sex and age of the mice are not relevant for this study.

Human tissues

Heparinized peripheral blood from one donor (10 - 40 mL) was obtained via venipuncture and processed using Ficoll density gradient centrifugation (Histopaque-1077, Sigma-Aldrich). The interface containing the peripheral blood mononuclear cells (PBMC) was extracted, washed twice with HBSS without Ca^{2+} and Mg^{2+} (Corning), and cryopreserved in FBS with 10% DMSO (10 - 20 millions/vial, 1mL). Human tonsils were obtained from the Stanford Tissue Bank (IRB protocol number 17204). The tonsils were cut into small pieces and the lymphocytes were released with HBSS containing Ca^{2+} and Mg^{2+} (Corning). The cell suspension was spun at 300 g, 4° C for 5 min; pelleted cells were frozen in FBS with 10% DMSO (10 - 20 millions/vial, 1mL). Samples were stored in liquid nitrogen until use. Sex, gender, and information about age of the samples are not relevant for this study.

Method details

Immunostaining for FACS and mass spectrometry

Flow cytometry of thymic T cells. Thymic T cells from two C57BL/6 male mice were immunophenotyped using a panel of antibodies to 13 markers involved in T cell development and maturation. Briefly, thymi were homogenized to obtain a single cell suspension. Cells were blocked with Fc-block (1:100) and rat serum (1:50) for 15 min at room temperature. We stained cellular surface in two steps: first primary antibody cocktail for 40 min, followed by 30 min secondary anti-biotin (streptavidin) stain (Key resource table). We fixed cells in 1% PFA for 5 min and stained intracellular content for Foxp3 in permeabilization buffer (eBiosciences) over night. Antibodies used in this study are indicated in the Key Resources Table.

On average, 100 000 events were collected using BD LSRFortessa™ (BD Biosciences) cell analyzer with 5 lasers. Some antibodies failed to stain significantly above background: these are not discussed further. Antibodies used for tSpace calculations are labeled in bold.

Mass spectrometry (CyTOF) antibodies. Mass cytometry antibodies were either purchased from Fluidigm (Sunnyvale, CA) or labeled in-house using MaxPAR X8 antibody conjugation kits (Fluidigm) following the manufacturer's protocol. Mass cytometry antibodies used are in Key resource table. Antibodies were diluted in Candor PBS Antibody Stabilization solution (Candor Bioscience GmbH, Wangen, Germany) supplemented with 0.02% NaN_3 (final concentration) and stored up to 6 months at 4° C.

Mass cytometry of human B cells in tonsils and blood. Cryopreserved PBMC or lymphocytes from tonsil were thawed at 37° C with gentle manual agitation, washed twice with HBSS without Ca^{2+} and Mg^{2+} supplemented with 10% bovine calf serum and resuspended at 20M cells/ml in CyFACS buffer (PBS with no heavy metal contaminants (Rockland Immunochemicals, Limerick, PA) with 1% FBS (Sigma) and

0.1% sodium azide). 5 million cells of each sample were used for staining. Cells were initially blocked for 10 minutes at 4° C with 0.5 µl normal human serum (Sigma) and 2.5 µl goat serum (Gibco) and then washed in CyFACS buffer (all subsequent washes were done with CyFACS buffer unless specified otherwise). Cells were incubated with purified, unconjugated anti-GPR15 for 45 minutes 4° C, washed once, then incubated with goat-anti-mouse IgG -168 for 30 minutes 4° C and washed. Cells were blocked with 2.5 µl normal mouse serum (Sigma) for 10 minutes 4° C, washed once and stained with a cocktail of primary antibodies for 45 minutes 4° C. Samples were washed and stained with metal-conjugated secondary antibodies against FITC to detect CLA-FITC and APC to detect $\alpha 4\beta 7$ -APC for 30 minutes 4° C and washed. Live/dead stain was performed for 30 minutes 4° C using ^{139}In -DOTA maleimide (Macrocyclics, Plano, TX) diluted 1:2000 in CyPBS (Rockland), washed twice and fixed overnight at 4° C in 2% paraformaldehyde (Electron Microscopy Systems). The next day, samples were washed twice and barcoded using the 20-plex Cell ID kit from Fluidigm according to the manufacturer's protocol except for two modifications. Cells were fixed overnight in 2% PFA instead of fixed in Fixl buffer (for 10 mins at RT) provided with the kit and barcoding was performed after surface staining rather than before. After barcoding, the samples were washed once in CyFACS, once in fixation/permeabilization buffer (eBiosciences) and then combined prior to intracellular staining. Intracellular staining was performed for 45 minutes 4° C using a cocktail of antibody conjugates against intracellular targets, washed twice with CyFACS buffer and DNA content stained for 20 minutes at room temperature using Ir-interchelator (Fluidigm) diluted 1:500 in 2% PFA. Final washes were performed (twice with CyFACS, once with CyPBS and once with milliQ water) then resuspended in 500 µl milliQ water.

The samples were filtered through cell strainer cap FACS tubes, counted and the cell concentration adjusted to 600 000 - 1 million cells/ml with milliQ H₂O containing bead standards (Fluidigm) prior to acquiring the sample on the CyTOF 2.0 mass cytometer (Fluidigm). Between 1 - 1.5 million total events were acquired. After the samples were manually de-barcoded by gating in FlowJo using the specific combination of three metals unique to each individual sample, 30 000 - 70 000 total cellular events remained.

Antibodies used in this study are indicated in the Key Resources Table. Cell markers used for tSpace calculation were: CCR6, CD19, IgD, CXCR5, CD20, CD69, CCR4, CD103, CLA, CCR2, CD22, CXCR3, CD77, CCR7, CCR9, CD62L, CCR10, CD27, GPR15, $\alpha 4\beta 7$, IgA, IgM, P-selectin, CD38 and IgG.

Quantification and statistical analysis

Computational processing of acquired data

Software. tSpace was performed in R [3.4.+] and MATLAB [9.3.0.713579 (R2017b)]. R language was used for other analyses: plots (plotly & shiny packages), t-SNE (Rtsne package) and UMAP (umap package), developmental branches isolation. All analysis code is available upon request. For gating of cell populations, we used FlowJo (10.2).

Processing and tSpace analysis of flow cytometry data. We inspected raw FCS files for artifacts and outliers in R or FlowJo and filtered out all events with fluorescence values higher or lower than 0.0001% quantile of the measured markers. We used t-SNE to filter out cells that were negative for all markers to be used in tSpace.

For thymus, NK1.1+ cells were removed manually in FlowJo. Remaining events were exported in a CSV file for tSpace. Thymic data were transformed using logicle (Parks et al., 2006) (included only in the MATLAB tSpace package). tSpace parameters were: distance metric = cosine, K = 15, L = 12, G = 5, WP = 20 and T = 1000. K-means clustering, as part of tSpace, on expression values of measured markers was used to define 1000 clusters for selection of the T (1000) trajectory start cells.

Processing of mass cytometry (CyTOF) FCS files. Data were normalized using the normalization software embedded within the CyTOF software. CyTOF records a value of zero for channels without detected metal. For better visualization, all zero values were assigned random values from a uniform distribution with minimum -1 and maximum 0 using a script in R. Data was transformed applying inverse hyperbolic sine (asinh function in R and MATLAB) with coefficient 5. We used t-SNE to find outliers, which

were gated out in FlowJo. For tonsil, CD19⁺ or CD38^{high} cells were gated for tSpace. Parameters in tSpace analysis were: distance metric = cosine, K = 17, L = 12, G = 5, WP = 20 and T = 1000 (Fig. 2, Fig. S6) or T = 100 (Fig. S5).

Processing of scRNAseq files. In intestinal dataset we used for tSpace calculation 2420 variable genes determined by the authors using Seurat R package (Butler et al., 2018). Parameters used for tSpace were distance metric = Pearson correlation, K = 20, L = 15, G = 5, WP = 15. We calculated ground truth trajectory matrix (for all cells, T = 3521).

For *C. elegans* dataset we selected all cells that passed original quality, normalized and scaled data as described (Packer et al., 2019) and performed tSpace analysis using 75 principal components in gene expression space with the following parameters: K = 25, L = 0.75*K, T = 1000, WP = 20, G = 5, distance_metric = pearson_correlation. Analysis was performed on a server with 48 cores and 254 GB memory.

tSpace running times. We measured the (reasonable) time required for tSpace analysis for 100 and 1000 trajectories (T) for our B cell tonsil dataset (17,956 cells, 26 variables), for the 3521 cell/2420 variable gene, 3521 cell/ 40 principal components scRNAseq sample and for our thymus dataset (~95,000 cells, 12 variables). Run times were obtained on a 2-core personal computer. *C. elegans* dataset (86,024 cells) was used to provide insight in analysis times of large datasets. We run all analysis on 48 core-computer with 254 GB memory. 50, 100, 500 and 1000 trajectories were calculated using 75 and 50 principal components from gene expression space.

Labeling of literature defined cell populations for tSpace validation. We manually gated on the indicated populations of T and B cells in FlowJo as illustrated in Fig. S3. and Fig. S4. Gated subsets were used to label and validate cell positions in tSpace visualization. For intestinal data we used cluster annotations from the original study for fully differentiated EE and enterocyte populations, however crypt associated (stem and transit amplifying) populations were labeled based on accepted markers along the developmental trajectories (Clevers, 2013) shown in Fig. S7. We used gene expression along the isolated developmental branches to identify and define crypt base columnar (CBC) cells, slowly cycling sc-CBC, cycling transit amplifying (c-TA) and TA cells. Examination of markers proposed to mark so-called Potten's "+4 cells" (Potten et al., 2003), suggested that these cells likely co-exist with or are the same as sc-CBC cells (Fig. S7C, E). Interestingly, many cells expressing transit amplifying markers and proliferation genes were assigned to already differentiated cells or cycling stem cells in the original publication (Fig. S7A-C). With tSpace we were able to detect, and position in putative developmental sequence, rare and transient populations (e.g. short lived enteroendocrine progenitors, sLEEP, Fig. 3D).

Isolation of developmental branches and calculation of expression changes along them. We manually gated on cells along specific developmental branches or pathways as visualized in the first 3-5 tPCs, using standard gating approaches in Flowjo, R or JMP. An example can be seen in vignette file accompanying our R package or online (<http://denisdermadi.com/tspace-trajectory-inference-algorithm>). Trajectories are directionless, thus, for orientation we relied on prior knowledge in the context of manually gated cell populations (e.g. thymic T and tonsil B cell developmental sequences) or expression of known marker of stem cells in intestine (*Lgr5*). Trajectory distances along the isolated developmental sequences were accessed from the tSpace distance matrices: we identified trajectories within the tSpace cell distance matrix that 'start' from cells at or near the putative origin of isolated sequences (i.e., trajectories for which the chosen start cells have the lowest distance value). Column names of trajectory space matrix contain cell indices associated with each trajectory in the distance matrix, facilitating identification of desired trajectories (see example online: <http://denisdermadi.com/tspace-trajectory-inference-algorithm>). One or more such trajectories were averaged for each developmental pathway isolated. Cells within isolated trajectories were ordered based on their trajectory distances and smoothed values of gene/protein expression along trajectories were visualized.

Systems biology analysis downstream of trajectory inference

Alignment of the intestinal trajectories and transcription factor analysis. Absorptive and secretory branches share many early cells e.g. CBC, sc-CBC and c-TA (Fig. 3A-C) before they separate fully in tSpace projection. In order to compare transcription factors expressed during initial branching of secretory and absorptive cells, early segments of secretory and absorptive trajectories were aligned using dynamic time warping (DTW) (Giorgino, 2009). DTW aligns sets of data points that can be represented as linear sequences, here cells ordered by tSpace along the absorptive and secretory branches. It calculates an optimal match between two sequences independently of rates of change. We performed DTW as a downstream analysis in R using dtw package and all variable genes to align the two branches, with absorptive as template and secretory as a query (Fig. S8B). DTW allowed us to split aligned sequences into 6 stages based on commonalities in gene expression. Significant differences in expression of mouse transcription factors (TFs, <http://tcofdb.org/>) within the same stages between the two trajectories were determined using a permutation test (Dermadi et al., 2014). Briefly, in permutation test we first define observed difference as a difference of medians of the same stage between the two trajectories (e.g. for each TF we calculate the difference between medians of TF expression for stage 1 of enterocyte and enteroendocrine trajectories). Then, TF expression values for each cell in stage 1 of these two trajectories are combined and randomly sampled to create two new groups (A and B) of sizes equal to the original groups. Then we calculate for each TF difference between medians of TF expression for group A and B. This is repeated 1000 times. Final statistics, p-value is defined as the sum of the absolute values of differences from permutation step that are higher or equal to the observed difference between the two trajectories and divided by the number of permutations. P-values are finally corrected for multiple testing using padjust R function and false discovery rate method.

TF modules were determined based on correlation of expression of significantly different TFs along the aligned trajectories. We define significant TFs with the p-value < 0.001. Highly correlated and anticorrelated TFs formed four modules visible in correlation matrix (Fig. S8C).

Gene ontology of transcription factors. Ontology terms of TFs were accessed using the R package Biomart and manually curated for biologically relevant terms highlighted in the Fig. S8. Many of the EE or enterocyte lineage TFs lack significant ontology annotation.

Data and code availability

Previously published dataset of intestinal cell populations was provided by the authors as normalized and scaled expression values (Yan et al., 2017). *C. elegans* dataset (86,024 cells) was downloaded from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE126954. Sources for code used in this study are indicated in the Key Resources Table.

Additional resources

An example and tutorial can be seen in vignette file accompanying our R package or online (<http://denisdermadi.com/tSPACE-trajectory-inference-algorithm>).

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|-------------------------|------------|
| Antibodies | | |
| CD90 | BD Biosciences | 53-2.1 |
| TCRb | BioLegend | H57-597 |
| CD3e | BD Biosciences | 17A2 |
| CD4 | BD Biosciences | RM4-5 |
| CD8a | eBioscience | 53-6.7 |
| CD24 | BD Biosciences | M1/69 |
| CD25 | BioLegend | PC61 |
| CD44 | eBioscience | IM7 |
| CD62L-biot | BioLegend | MEL-14 |
| CD117 | BD Biosciences | 2B8 |
| CCR9 | BioLegend | CW-1.2 |
| NK1.1 | BioLegend | PK136 |
| Streptavidin | BD Biosciences | |
| Foxp3 | eBioscience | FJK-16s |
| CCR6 | Fluidigm | G034E3 |
| CD19 | Fluidigm | HIB19 |
| IgD | Biolegend | IA6-2 |
| CXCR5 | Biolegend | RF8B2 |
| CD4 | Fluidigm | RPA-T4 |
| CD8 | Biolegend | SK1 |
| CD20 | Fluidigm | 2H7 |
| CD69 | Biolegend | FN50 |
| CCR4 | Fluidigm | 205410 |
| CD3 | Biolegend | UCHT1 |
| CD103 | Biolegend | BerAct8 |
| TCRgd | Fluidigm | 11F2 |
| CLA | Biolegend | HECA452 |
| CCR2 | Biolegend | K036C2 |
| CD22 | Biolegend | HIB22 |
| CXCR3 | Fluidigm | G02SH7 |
| CD14 | Biolegend | M5E2 |
| CD33 | Fluidigm | WM53 |
| CD77 | Biolegend | 5B5 |
| CCR7 | Biolegend | G043H7 |
| Ki67 | Fluidigm | B56 |
| CCR9 | Produced from Hybridoma | L053E8 |
| CD62L | Produced from Hybridoma | Dreg200 |
| CCR10 ² | Produced from Hybridoma | 1B5 |
| CD27 | Fluidigm | L128 |
| GPR15 | R&D Systems | 367902 |
| goat-anti-mouse IgG | Invitrogen | 10535 |
| a4b7-APC | Produced from Hybridoma | Act-1 |
| Anti-APC | Biolegend | APC003 |
| IgA | BD Biosciences | Polyclonal |
| b7 | Biolegend | FIB504 |
| IgM | Fluidigm | MHM-88 |

| | | |
|--|---|---|
| P-selectin Ig CD38 CD56 IgG CD16 | R&D Systems Biolegend Biolegend BD Biosciences Fluidigm | 137-PS HIT2 HCD56 Polyclonal 3G8 |
| Biological Samples | | |
| Cryopreserved PBMC or lymphocytes from tonsil | Stanford Tissue Bank | N/A |
| C57BL/6J male mice | VA Animal Facility | N/A |
| Chemicals, Peptides, and Recombinant Proteins | | |
| HBSS without Ca ²⁺ and Mg ²⁺ | Corning | 21-021-CV |
| Fetal Bovine Serum (FBS) | Sigma-Aldrich | ES-009-C |
| Dimethyl sulfoxide (DMSO) | Sigma-Aldrich | D2650 |
| HBSS with Ca ²⁺ and Mg ²⁺ | Corning | 21-020-CV |
| Fc-block | ThermoFisher - Invitrogen | 14-9161-73 |
| Rat serum | Sigma-Aldrich | S24-M |
| Permeabilization Buffer (10X) | ThermoFisher - eBiosciences | 00-8333-56 |
| MaxPAR X8 antibody conjugation kits | Fluidigm | Catalog numbers depending on the metals: 201141A–201156A and 201158A–201176A |
| Candor PBS Antibody Stabilization solution | Candor Bioscience GmbH, Wangen, Germany | 130 050 |
| Sodium azide (NaN ₃) | Sigma-Aldrich | S8032 |
| Phosphate Buffered Saline (PBS) | Rockland Immunochemicals, Limerick, PA | #MB-011 |
| Normal human serum | Sigma-Aldrich | S1-100ML |
| Goat serum | ThermoFisher - Gibco | 16210064 |

| | | |
|--|---|---|
| Normal mouse serum | Sigma-Aldrich | NS03L-1ML |
| Maleimido-mono-amide-DOTA | Macrocyclics, Plano, TX | B-272 |
| Paraformaldehyde | Electron Microscopy Systems | RT 15700 |
| Cell ID kit | Fluidigm | 201060 |
| Ir-interchelator | Fluidigm | 201192B |
| EQ™ Four Element Calibration Beads, bead standards | Fluidigm | 201078 |
| Histopaque-1077 | Sigma-Aldrich | 10771-100ML |
| Deposited Data | | |
| single cell data provided by the authors | Yan et al., 2017 | GSE99457 |
| <i>C. elegans</i> single cell data | Packer et al., 2019 | GSE126954 |
| Software and Algorithms | | |
| tSpace package R version | this study | https://github.com/hylasD/tSpace |
| tSpace MATLAB version | this study | https://github.com/hylasD/MATLAB_version_tSpace |
| Seurat (v2+) | Butler et al. 2018; | https://satijalab.org/seurat/ |
| R (3.4+) | N/A | https://www.r-project.org/ |
| Rstudio | N/A | https://www.rstudio.com/ |
| MATLAB (2017) | N/A | https://www.mathworks.com/products/matlab.html |
| FlowJo (v10.2) | N/A | https://www.flowjo.com/ |
| ggplot2 | N/A | http://ggplot2.org/ |
| PAGA | doi.org/10.1101/208819 | https://github.com/theislab/paga |

| | | |
|--------------------------|---|---|
| destiny | 10.1038/nmeth.3971 | https://bioconductor.org/packages/release/bioc/html/destiny.html |
| p-creode | 10.1016/j.cels.2017.10.012 | https://github.com/KenLauLab/pCreode |
| Monocle | 10.1038/nmeth.4402 | http://cole-trapnell-lab.github.io/monocle-release/ |
| umap | https://arxiv.org/abs/1802.03426 | https://cran.r-project.org/web/packages/umap/index.html |
| Rtsne | http://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf | https://cran.r-project.org/web/packages/Rtsne/index.html |
| plotly | N/A | https://plot.ly/r/ |
| shiny | N/A | https://shiny.rstudio.com |
| dtw | doi:10.18637/jss.v031.i07 | https://dynamictimewarping.github.io |
| Biomart | 10.18129/B9.bioc.biomaRt | https://bioconductor.org/packages/release/bioc/html/biomaRt.html |
| Other | | |
| CyTOF 2.0 mass cytometer | Fluidigm | |
| tSpace tutorial | denisdermadi.com | http://denisdermadi.com/t-space-trajectory-inference-algorithm |
| BD LSRT Fortessa™ | BD Biosciences | |

Supplemental References

Bendall, S.C., Davis, K.L., Amir, E., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., Pe'er, D., 2014. Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development. *Cell* 157, 714–725. <https://doi.org/10.1016/j.cell.2014.04.005>

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36, 411. <https://doi.org/10.1038/nbt.4096>

Clevers, H., 2013. The Intestinal Crypt, A Prototype Stem Cell Compartment. *Cell* 154, 274–284. <https://doi.org/10.1016/j.cell.2013.07.004>

Đermadi, D., Valo, S., Pussila, M., Reyhani, N., Sarantaus, L., Lalowski, M., Baumann, M., Nyström, M., 2014. Inherited cancer predisposition sensitizes colonic mucosa to address Western diet effects and putative cancer-predisposing changes on mouse proteome. *J Nutritional Biochem* 25, 1196–1206. <https://doi.org/10.1016/j.jnutbio.2014.06.002>

Dijkstra, E., 1959. A note on two problems in connexion with graphs. *Numer Math* 1, 269–271. <https://doi.org/10.1007/bf01386390>

Giorgino, T., 2009. Computing and Visualizing Dynamic Time Warping Alignments in R : The dtw Package. *Journal of Statistical Software* 31. <https://doi.org/10.18637/jss.v031.i07>

Haghverdi, L., Büttner, M., Wolf, A.F., Buettner, F., Theis, F.J., 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* 13, nmeth.3971. <https://doi.org/10.1038/nmeth.3971>

Parks, D.R., Roederer, M., Moore, W.A., 2006. A new “Logicle” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytom Part A* 69A, 541–551. <https://doi.org/10.1002/cyto.a.20258>

Potten, C.S., Booth, C., Tudor, G.L., Booth, D., Brady, G., Hurley, P., Ashton, G., Clarke, R., Sakakibara, S., Okano, H., 2003. Identification of a putative intestinal stem cell and early lineage marker; musashi-1. *Differentiation* 71, 28–41. <https://doi.org/10.1046/j.1432-0436.2003.700603.x>

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 2319–2323. <https://doi.org/10.1126/science.290.5500.2319>

Yan, K.S., Gevaert, O., Zheng, G., Anchang, B., Probert, C.S., Larkin, K.A., Davies, P.S., Cheng, Z., Kaddis, J.S., Han, A., Roelf, K., Calderon, R.I., Cynn, E., Hu, X., Mandleywala, K., Wilhelmy, J., Grimes, S.M., Corney, D.C., Boutet, S.C., Terry, J.M., Belgrader, P., Ziraldo, S.B., Mikkelsen, T.S., Wang, F., von Furstenberg, R.J., Smith, N.R., Chandrakesan, P., May, R., Chrissy, M.S., Jain, R., Cartwright, C.A., Niland, J.C., Hong, Y.-K., Carrington, J., Breault, D.T., Epstein, J., Houchen, C.W., Lynch, J.P., Martin, M.G., Plevritis, S.K., Curtis, C., Ji, H.P., Li, L., Henning, S.J., Wong, M.H., Kuo, C.J., 2017. Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Cell Stem Cell* 21, 78-90.e6. <https://doi.org/10.1016/j.stem.2017.06.014>