OXFORD

TIPS: trajectory inference of pathway significance through pseudotime comparison for functional assessment of single-cell RNAseq data

Zihan Zheng, Xin Qiu, Haiyang Wu, Ling Chang, Xiangyu Tang, Liyun Zou, Jingyi Li, Yuzhang Wu, Jianzhi Zhou, Shan Jiang, Ying Wan and Qingshan Ni

Corresponding authors: Shan Jiang, Institute for Advanced Study, Shenzhen University, Shenzhen, China. E-mail: shan.jiang1975@szu.edu.cn; Ying Wan; Qingshan Ni, Biomedical Analysis Center, Army Medical University, Chongqing, China. E-mail: wanying516@foxmail.com; qingshanni@foxmail.com

Abstract

Recent advances in bioinformatics analyses have led to the development of novel tools enabling the capture and trajectory mapping of single-cell RNA sequencing (scRNAseq) data. However, there is a lack of methods to assess the contributions of biological pathways and transcription factors to an overall developmental trajectory mapped from scRNAseq data. In this manuscript, we present a simplified approach for trajectory inference of pathway significance (TIPS) that leverages existing knowledgebases of functional pathways and other gene lists to provide further mechanistic insights into a biological process. TIPS identifies key pathways which contribute to a process of interest, as well as the individual genes that best reflect these changes. TIPS also provides insight into the relative timing of pathway changes, as well as a suite of visualizations to enable simplified data interpretation of scRNAseq libraries generated using a wide range of techniques. The TIPS package can be run through either a web server or downloaded as a user-friendly GUI run in R, and may serve as a useful tool to help biologists perform deeper functional analyses and visualization of their single-cell data.

Key words: trajectory mapping; pseudotime; TIPS; pathway analysis

Zihan Zheng is a bioinformatician working on biological applications of trajectory analysis.

Xin Qiu is a technician specializing in performing and optimizing methods for single-cell library construction.

Haiyang Wu is a bioinformatician working on identifying antigen-specific T cell clones through single-cell sequencing.

Ling Chang is a PhD student at Army Medical University working on applying scRNAseq to interrogate T cell differentiation and clonal behavior in autoimmune diseases.

Xiangyu Tang is a PhD student at Army Medical University working on applying scRNAseq to identify new mechanisms involved in osteoclast differentiation.

Liyun Zou is an Assistant Professor at Army Medical University working in the fields of antigen presentation and autoimmunity.

Jingyi Li is a Clinician and an Assistant Professor in the Department of Rheumatology and Immunology who specializes in autoimmune diseases and has projects focused on exploring the impact of pharmaceutical intervention on immune response.

Yuzhang Wu is the Director of the Immunology Institute at Army Medical University who specializes in T cell immunology during autoimmunity and chronic infection.

Jianzhi Zhou is the Research Director at Biowavelet Ltd., a company working on developing hardware and software for automation of high-throughput sequencing library construction.

Shan Jiang is an Assistant Professor at Shenzhen University working in the field of T cell development and differentiation.

Ying Wan is the Director of the Biomedical Analysis Center at Army Medical University and has led work focused on single cell sequencing and vesicular transport.

Qingshan Ni is a bioinformatician and lecturer at Army Medical University whose main research focus is on data analysis for high-throughput sequencing.

Introduction

Recent advances in single cell RNAseq (scRNAseq) library construction technologies, coupled with decreases in cost of high-throughput sequencing, have led to a proliferation of new single cell transcriptome data across a range of species and developmental/disease contexts. In order to interpret this data, a slew of new bioinformatics tools and analysis techniques have been developed. These include network-based methods to infer *cis*-regulatory interactions, splicing-informed mapping of RNA-velocity and inferences on key intercellular interactions based on receptor-ligand pairs [1–3]. These and other novel tools have helped to unlock the potential of scRNAseq to provide a plethora of information that was previously inaccessible to analysis of bulk sequencing-based datasets.

However, one area in which currently available single-cell analysis tools are lacking in is in pathway analysis. While pathways analysis and similar gene-set-based enrichment analyses are among the most common ways to infer molecular mechanisms contributing to altered behavior under different conditions, directly applying bulk-sequencing inspired pathway analysis methods to scRNAseq data comes with substantial difficulties. Unlike bulk sequencing profiles, scRNAseq transcriptome tend to have significant dropout, with technical variation causing false-zero detection of any given gene in a cell [4, 5]. Chance-driven technical dropout also tends to be particularly severe for genes with low- to medium-expression magnitudes, many of which are key components of biological pathways. Because of this, some approaches (such as Metacell) have been developed to enable gene set enrichment analyses across clusters of cells using composite/imputed profiles [6]. However, such an approach sacrifices the single-cell resolution offered by scRNAseq by using it to approximate bulk sequencing results. Furthermore, this emphasis on pairwise differential regulation is less suitable for single cell datasets that encompass multiple cellular states and functional clusters. Other tools, such as AUCell and GSVA, preserve single-cell resolution to generate scored profiles for each cell for a given pathway but are focused on describing these pathways in terms of absolute expression level [7, 8]. However, not all pathway changes can be accurately reflected as absolute increases in expression level; many curated pathways will also contain genes with weak changes or decreases in expression as part of known regulatory behavior. In addition, these approaches cannot provide information on the temporal order by which pathways (and the genes represented within) may change during a larger biological process.

In order to implement pathway analysis at single-cell resolution, we present here a novel analytical framework that provides trajectory inference of pathway significance (TIPS). Our approach leverages the trajectory mapping principle of pseudo-time assignment to build pathway-specific trajectories from a pool of single cells. The pseudotime values for each cell along these pathway-specific trajectories are then compared to identify the processes with highest similarity to an overall trajectory in a simple and intuitive process. This approach differs from existing pathway analysis methodologies by evaluating expression dynamics of individual genes, instead of taking an averaged measure of gross expression. Consequently, this approach allows for straightforward comparisons of the relative influence of each pathway, as well as the timing of pathway changes. Furthermore, key genes that are associated with both the overall trajectory and/or pathway-specific trajectories are also identified, providing ready targets for downstream validation work. Direct visualizations are also offered at each of the primary steps, with

customizable options for figure generation for pathways and genes of interest. The key modules in TIPS incorporate in a number of leading scRNAseq analysis tools and can be run sequentially within a shiny GUI in R with the source code available on GitHub (https://github.com/qingshanni/TIPS) or on a dedicated webserver (http://118.24.236.198:3838/TIPS/). We hope that the TIPS workflow may help further expand the range of functional analyses that are possible when working with scRNAseq data, and help derive new functional insights into complex biological processes.

Materials and Methods

Overview of the TIPS framework

The TIPS framework is designed as 7 primary modules (described in detail below) intended to be run sequentially in R as either a local shiny-based GUI or as online as a webserver. Each module will automatically generate and format the analysis results as necessary to permit further analyses without a need for manual modification. A schematic overview of these modules is included as Figure 1, and a video walkthrough of the operation of the shiny-based GUI using demo data is included as Supplementary Video S1, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Uploading data

TIPS takes as its primary data input a standard gene expression matrix, wherein each row corresponds to a gene and each column a cell, from a comma-delimited (.csv) file. Ideally, this matrix should be pre-processed beforehand to only include in the cells that are of sufficient quality and which are pertinent to the intended trajectory analysis. Since preprocessed data may have been pre-normalized, an option is available to either log-normalize raw data or to accept it as is. TIPS also accepts metadata information for each cell (regarding sample quality, origin, type, etc.), from an additional (.csv) file for further visualizations. The user may then choose a list of gene sets to consider from a dropdown menu (six are built-in, including Reactome, KEGG, BioCarta, Msigdb) or otherwise upload an additional (.gmt) file containing the gene lists they wish to analyze [9–11]. This allowance for built-in options allows for customized analysis of data derived from other organisms or knowledgebases.

Dimension reduction

TIPS automatically loads and manages the data input as a Seurat object and runs three different dimension reduction algorithms (PCA, tSNE and UMAP) to generate 2D visualizations of the transcriptome similarity between cells [12, 13]. Users may select cutoff parameters for selecting the list of highly variable genes (HVGs) to consider for these reductions (both expression and dispersion cutoffs), and number of nearest neighbors, in order to optimize the analysis to suit their dataset. The results from the reduction can be visualized directly, with metadata information as selectable overlays, while a Louvain clustering algorithm is run at common resolutions to provide a range of clustering results.

Primary trajectory analysis

As a matter of principle, the TIPS framework can next be applied directly to generate a primary trajectory that includes all cells.
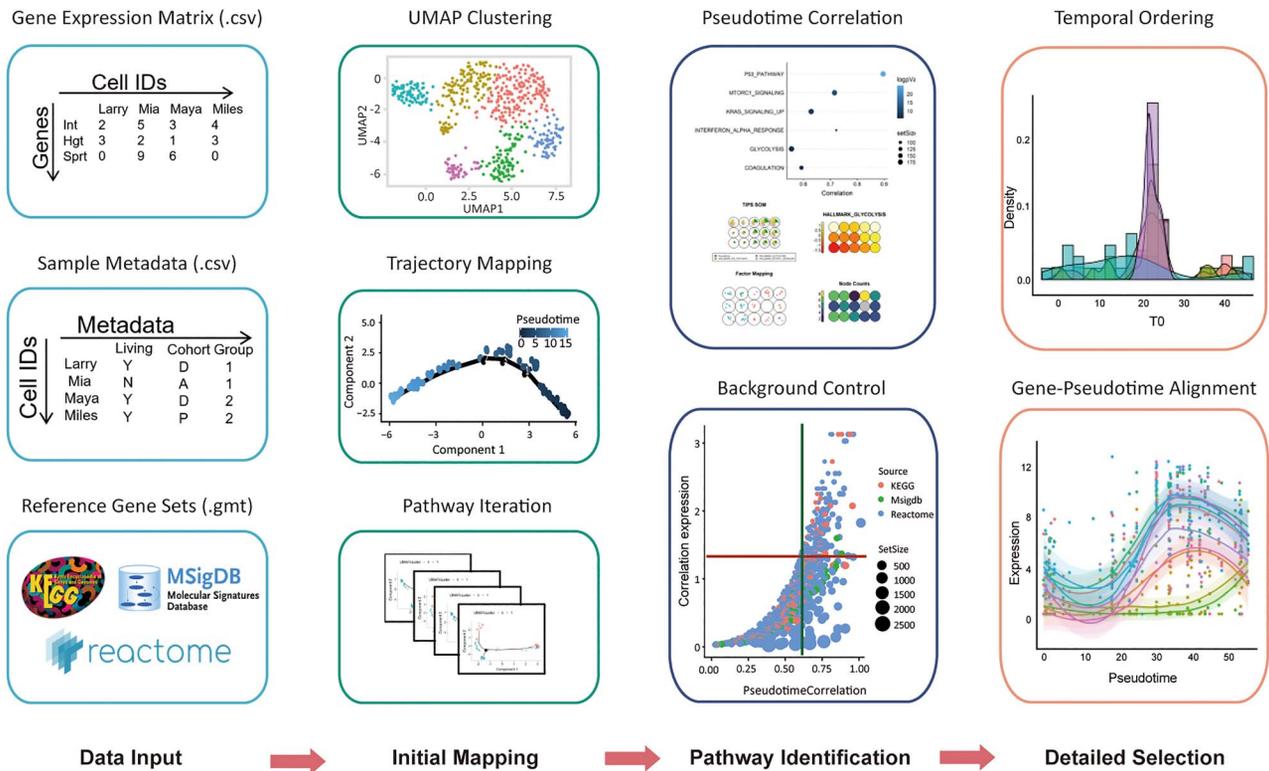
Figure 1. Overview of TIPS workflow. The TIPS workflow as presented here and as implemented in our GUI and shiny app requires the input of only a gene expression file and the selection of a reference database for consideration, with an optional option for helping visualize associated metadata. These data are then loaded to create a Seurat object. Three modes of dimension reduction are then run (PCA, tSNE and UMAP) to generate 2D visualizations of transcriptional similarity between cells using HVGs. A trajectory is constructed using DDRTree in monocle (run in successive order in our app). Users may then select parameters regarding gene size and expression to filter for pathways of interest, and additional trajectories are then constructed based on individual pathway gene sets. The statistical significance of each pathway is computed based on signal relative to background noise from 1000 randomly selected gene lists of the same size. Users may then further select specific pathways of interest for further analysis. These analyses include temporal ordering of different pathways, as well as selection of the critical genes from each pathway for downstream validation. All visualizations can be exported as publication-ready PDFs or tiffs with selectable scaling. The Seurat and monocle objects are stored in a directory to allow for further manipulation by the user if necessary. Full tables of pathways and genes within displaying significant association are also provided as text files.

However, we strongly recommend that users perform any data subsetting or sample filtering steps before the trajectory analysis step to ensure that the analysis can return meaningful results in practice. In particular, we suggest that users only select cell types/clusters which have direct temporal relations and/or transitional potential for downstream analysis in TIPS. This is because inclusion of irrelevant cell types will significantly increase the calculation runtime at the expense of computational accuracy. Inclusion of extraneous clusters introduces noise to the analysis, which can lead to false positive inferences for pathways that are not related to the primary trajectory of interest, while also sacrificing the resolution of pathways with actual trajectory correlation. As such, users of the TIPS GUI are given the option to select clusters of interest identified by the preceding dimension reduction visualization, with access to the heatmap visualization of the top markers for each cluster.

After selecting the clusters of interest, the HVGs that define these clusters are then calculated, to ensure representational accuracy. These HVGs are then passed to Monocle for DDRTree-based mapping and pseudotime assignment to generate the primary dataset trajectory [14]. Since the actual assignment direction for pseudotime values can be arbitrary, an option to reverse the initial order of cells is provided, such that users may select an order more intuitive for their question of interest in terms of expected temporal relations between clusters.

This assignment is then treated as the reference pseudotime trajectory for downstream analysis.

Pathway inference

In order to identify the pathways that are functionally associated with the overall trajectory of interest, we first load into the TIPS pipeline the selected reference knowledgebase of pathways for consideration. We then tailor these reference pathways to the dataset of interest to remove genes with missing expression and prune out overly small pathways based on absolute size (<20 genes). TIPS then iteratively runs the DDRTree algorithm to generate one trajectory per pathway of interest, considering all of the expressed genes from the pathway (including ones where the genes display non-significant variation in expression) to give a complete picture of pathway dynamics. This process reduces each pathway to a single linear vector of pseudotime values. These pseudotime vectors are then compared with the reference pseudotime trajectory using absolute Pearson's correlation to assess relative similarity in terms. Notably, absolute comparisons are used for these measures to avoid potentially misleading inferences about the direction of trajectory values. In order to assess the pathways with true significance, DDRTree is also iterated on randomly generated pathway lists to establish a false discovery rate (FDR) for genes sets of a given size. Pathways

at correlation levels below 5% FDR and which have significant correlation in terms of absolute gene expression (>0.6) are considered to be significantly associated with the overall trajectory.

Pathway temporal alignment

Individual significantly associated pathways may have distinct modes of behavior over the course of the pseudotime trajectory, with some showing rapid changes early on, while others are late-breaking. In order to provide a temporal understanding of these pathway dynamics, we further utilized the switchde package in R to help identify genes with switch-like expression characteristics [15]. We can then visualize the distribution of switch points with respect to pseudotime for each individual pathway and subsequently compare these distributions to generate a temporal alignment.

SOM-based pathway selection

Although the absolute order of pseudotime values may be arbitrarily assigned and is not dependent upon a unidirectional change in expression within a given gene list, researchers may be interested in focusing on investigating processes that do show continual increases or decreases expression over time. In order to accommodate this need, we implemented self-organizing maps (SOMs) to cluster the cells using the kohonen package in R [16]. Users can select their pathways of interest from a dropdown menu to assess the direction while also choosing the number of nodes for the SOM that would best fit their interest. Pathways that display monotonic changes in behavior can thus be visualized in this manner. The SOM can also be used to help visually distinguish pathways with more complex expression kinetics.

Individual gene selection

Although the approach outlined above is sufficient for identifying pathways that meaningfully contribute to an overall trajectory, it is oftentimes necessary to further clarify the exact genes that drive such a contribution. To address this question, we included three distinct methods for identifying critical genes within a pathway. First, we consider the genes with switch-like behavior recovered from switchde, as the singular and abrupt change in expression within these genes make them good candidates for functional screening and validation. Second, since other critical genes may display milder and monotonic changes in expression, we also compute the Pearson correlation between gene expression and pseudotime progression to help identify genes not found by switchde. Finally, since still other factors may display more complex changes in expression dynamics (alternating increases and decreases, or the like), we further weigh the relative contribution of a given gene on the pseudotime correlation of its parent pathway. Through these three distinct approaches contained in our final module, we can subsequently narrow down the range of candidate genes to consider for further validation work.

Analyzed datasets

For our initial analysis of a simulated dataset, we used the splatter package in R with default settings to generate a dataset of 500 cells with 5000 expressed genes per cell via the simulating paths option [17]. To enable analysis of the impact of technical dropout, we further used the dropout function in splatter to add in zero-inflation up through a range of median expression values. To clarify that these conclusions were also applicable to larger and more complex datasets, we further simulated datasets with 1, 2 and 3 distinct paths and larger numbers of cells, as well as a dense dataset of 5000 cells and a pool of 15 000 genes.

scRNAseq library construction and sequencing

In order to provide a direct demonstration of the use of the TIPS framework on real scRNAseq data, we isolated peripheral blood mononuclear cells from whole blood provided by a healthy donor under approval of the Ethics Committee of Southwest Hospital as part of a pilot study. CD8+ T cells were acquired and FACS-sorted (Beckmann Coulter) into a 96-well plate following staining with antibodies against CD3, CD4 and CD8 (BD). Cells were then lysed, and libraries were prepared using the scSTATseq workflow that we have previously developed [18]. Libraries were sequenced using the HiSeq 4500 platform (Illumina) and preprocessed as previously described. Alignment-free counting of reads relative to the reference human transcriptome was performed using Salmon [19]. The resulting gene expression matrix was then read into the Seurat package in R, where quality control filtering was performed to remove cells with excess mitochondrial reads (>10% of all reads), as well as outliers in terms of number of genes recovered (<5000 or >15 000). The remaining 69 libraries were then passed to the TIPS framework for further analysis and visualization.

Additional comparison datasets

Additional data from scRNAseq libraries constructed using alternative methods were obtained from GSE133535, a benchmarking study of different library construction protocols [20]. Since a mix of cells were used in the study, we individually downloaded each of the data matrices corresponding to a single method and filtered them based on metadata annotation to only retain human cells. Data on these cells were then loaded into Seurat for UMAP-based clustering in order to identify clusters of CD8+ T cells (based on positive expression of CD8, CD3D and lack of expression of CD4 and NCAM1). Transcriptome information from these CD8+ T cells was then passed to TIPS for additional analysis.

For analysis of single CD8+ T cells from HCC patients, fully processed sparse matrix files were obtained from GSE98683, and annotated CD8+ populations were used for further analysis through TIPS [21]. Data for the HCC samples treated with immune checkpoint blockade (ICB) were similarly processed from GSE125449 [22]. For our analysis of endocrine cells from the pancreatic islets of patients with type II diabetes, data were obtained from GSE86469 [23]. A summary of the gene expression characteristics of these datasets is included as Supplementary Figure S11, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Results

Testing of the TIPS workflow using simulated data

In order to assess the robustness of the analysis workflow described, we first simulated a medium-sized scRNAseq dataset of 5000 genes and 500 cells ordered along a single path. We then compared the performance of different correlation metrics for comparing pathways to a common trajectory using 1000 randomly selected gene lists and 100 HVG lists. Since the baseline

pseudotime trajectory was generated based on a complete list of HVGs, we anticipated that subsets of the HVG list would show true signal, while the completely random lists would reflect the range of noise. Interestingly, we observed that the HVG lists were essentially indistinguishable from random background when mean pathway gene expression level was considered (Figure 2A). However, pseudotime-to-pseudotime correlations showed a clear separation between HVG lists and background (Figure 2B). This phenomenon suggested that the TIPS workflow will highlight pathways as being highly significant regardless of average expression (Figure 2C). Furthermore, this pattern held when datasets of larger size and containing multiple independent paths were considered, suggesting that it is highly robust (Supplementary Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). Instead, we observed that the constituent genes from the highly correlated lists displayed balanced distribution, with a similar number of genes having increasing or decreasing expression (and changes of similar magnitude) over the course of the pseudotime trajectory (Figure 2D). This trend could also be clearly identified in terms of specific genes from a single pathway (Figure 2E). At the same time, we also observed a clear pattern wherein the addition of additional information from more HVGs led to higher correlation with the overall pseudotime (Figure 2F). As such, these results suggest that pseudotime correlation provides a sensitive method for comparing trajectories, by accounting for genes with both increasing and decreasing expression.

At the same time, we also assessed the potential influence of other factors on the accuracy of pseudotime mapping to clarify its range of applicability. Since scRNAseq libraries may vary greatly in terms of number of cells recovered, genes recovered and in sequencing quality, we controlled for each of these factors in turn. Interestingly, pseudotime mapping using subsampled cells demonstrated that decreasing the size of the dataset did not have a strong of an effect on correlation accuracy, such that a correlation above 0.9 could be maintained even when the dataset was downsampled to 10% of its original size (Figure 2G). This trend held when we repeated this analysis in a much larger dataset of 5000 cells and 15 000 genes (Supplementary Figure S2, see Supplementary Data available online at http://bib.oxfordjournals.org/). This result suggests that sequencing a smaller number of representative cells may be sufficient to map a trajectory. Since real scRNAseq data often face practical limitations regarding the numbers of cells recovered from a population of interest and can include multiple paths, these results show that the accuracy trajectory mapping will not be constrained by these limitations.

However, the gene detection quality of the cells sequenced had a substantial impact on accuracy, as increasing the degree of technical dropout in the data led to a clear deterioration in the correlation (Figure 2H). This deterioration demonstrates that pseudotime assignment is sensitive to information loss. In a similar vein, we observed that the number of genes considered for pseudotime assignment also had significant influence on its accuracy; larger sets of randomly selected genes tended to have significantly higher background noise than their smaller counterparts (Figure 2I). Since a given dataset may only feature significant expression in a subset of all genes assigned to a given pathway, the relative size and representation rate of a pathway may also influence correlation interpretation. As such, we elected to run independent calculations of background noise using random lists for each pathway in order to control for differences in gene list size and representation. These latter two limitations are important caveats to consider for trajectory analysis as a whole and are not unique to using the TIPS framework. However, since TIPS relies on iterating the trajectory mapping process, TIPS is unlikely to infer many significant pathways from datasets with relatively poor quality in these two metrics.

TIPS confirms existing knowledge of CD8+ T cell differentiation

To validate the utility and biological relevance of our workflow, we then analyzed 69 single-cell libraries of peripheral blood CD8+ T cells from a healthy donor using the scSTAT-seq method. UMAP clustering of the cells readily identified two prominent clusters of cells of similar size (Figure 2A), and marker analysis demonstrated that cluster0 was composed of antigen-experienced effector cells positive for the effector molecules IFNG and GZMB, while cluster1 included naïve/memory cells that displayed high levels of CCR7 and S1PR1 (Figure 2B). Pseudotime trajectory mapping based on the dispersed genes yielded a relatively simple arc, with the effector cells being assigned higher pseudotime values along this reference trajectory (Figure 2C). We then generated iterated trajectory mappings using the curated hallmark signatures from three separate knowledgebases (Msigdb, KEGG, Reactome) and performed paired correlation analysis of the pseudotime values against the reference trajectory to identify pathways with close association and significant signal over noise based on gene set size (Figure 2D). Consistent with our expectations, a relatively small portion of pathways showed significance, alleviating concerns about overfitting from our method (Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). Of note, no correlation was found between the number or percentage of HVGs in a given pathway and its inferred significance (Supplementary Figure S3A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/), demonstrating that the results obtainable from this approach are distinct from enrichment of the HVGs list. Instead, we observe that even pathways without an HVG can still be inferred to be functionally important, as a result of weaker, but consistent, changes in expression across large numbers of genes, while some lists containing large numbers of HVGs may still have insignificant contribution (Supplementary Figure S3C and D, see Supplementary Data available online at http://bib.oxfordjournals.org/).

From direct inspection of the top pathways displaying significant association via TIPS analysis, we found a number of well-characterized processes known to influence CD8 behavior, such as chemokine signaling and IL12 family signaling (Figure 2E). At the same time, we also observed significance in less appreciated processes such as the Myd88-mediated, TLR-cascaded and SLC-mediated transmembrane transport. Overall temporal alignment of these four pathways of interest based on the order of their switched-on genes demonstrated that each of these pathways had factors that changed across multiple points of the overall pseudotime trajectory, although most of the changes were centered at an intermediate timepoint marking a changeover from memory to effector status (Figure 2F). When we focused in on each pathway, we could observe that these changes included relatively critical constituents, such as the signaling kinase IRAK1 in the Myd88-mediated cascade. Many of these molecules have also previously been demonstrated to be of functional importance in the context of T cell immunity, such as the co-stimulatory molecule ICOS, the large neutral amino acid transporter SLC7A5 and the chemokine receptor CXCR1 [24–26]. We also recovered several
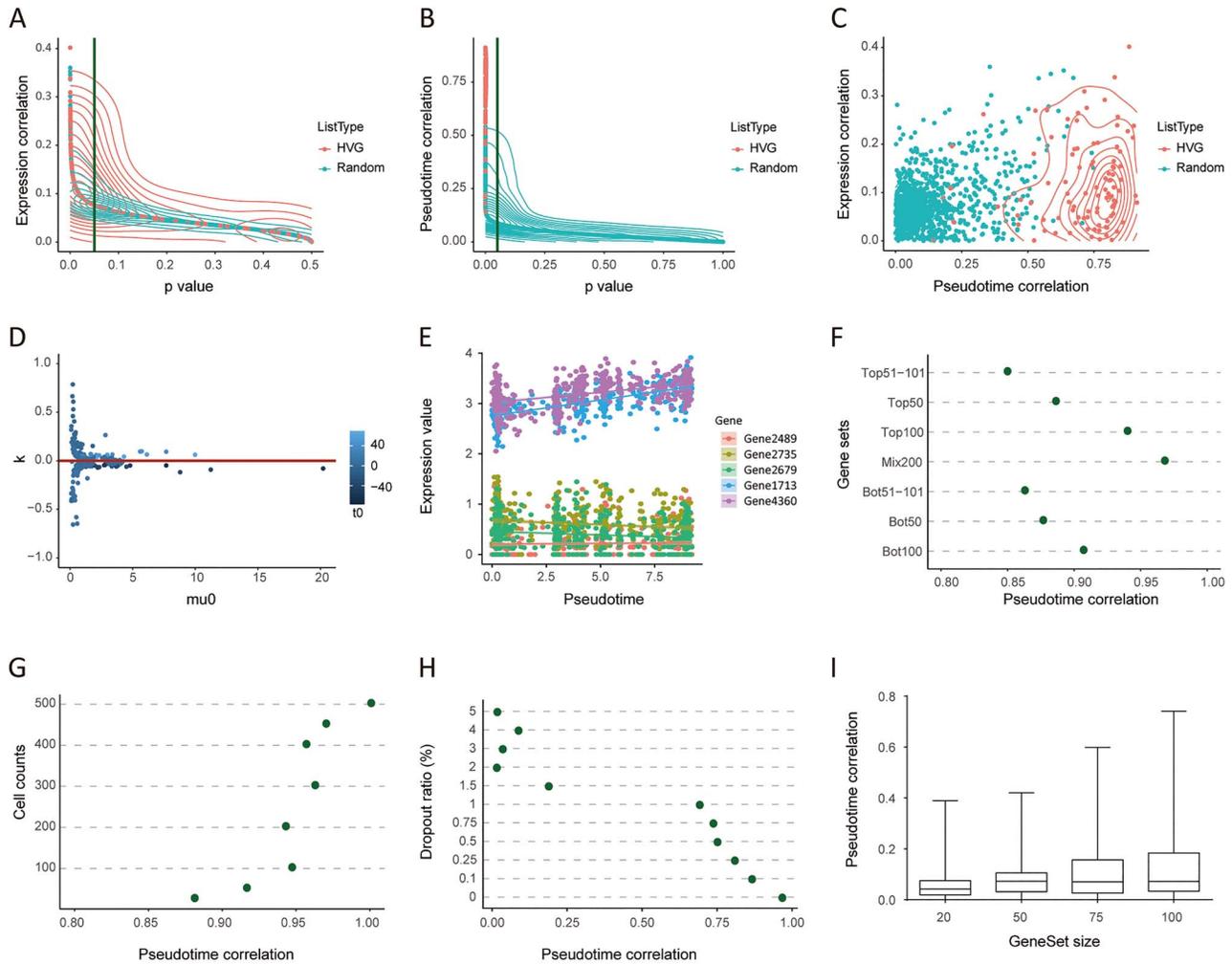
Figure 2. Parameter testing using simulated data for parameter testing, we used a simulated dataset of 500 cells and 5000 genes per cell, comparable in information amount to most real single-cell datasets. We then generated 1000 lists of 100 genes each for background measurement and 100 lists of 100 genes each from the subset of HVGs (527 total) as our signal of interest. (A) Average gene expression is commonly used as a metric for identifying significantly associated pathways in bulk analyses. (B) When the lists were instead run through DDRTree to generate individual pseudotime vectors, pseudotime vectors derived from HVG lists showed very high levels of correlation, while few random lists showed significance. P-values shown in A and B are for the Pearson correlation statistic. (C) Dotplot of pseudotime correlation versus expression correlation demonstrates that the majority of HVG lists that showed high levels of pseudotime correlation had very little correlation in terms of expression ($R < 0.2$). (D) Dotplot of the expression profiles of all HVGs of the simulated dataset derived from switchde. mu0 indicates the half-peak expression of a gene prior to the switch event, while k indicates the magnitude of the switch event, and t0 the timepoint along the trajectory at which the switch event takes place. (E) Taking one significantly associated HVG list as an example, we can observe that a given pathway may also include individual genes with increasing and decreasing expression. (F) By increasing or decreasing the size of the gene list considered, we can observe that the magnitude of the pseudotime correlation is highly sensitive to the size and amount of information considered. These results confirm that the method is sufficiently sensitive. (G) Real scRNAseq datasets can vary greatly in the numbers of cells sequenced. By subsetting the simulated set, we sought to measure the influence of changes in dataset size. Interestingly, while the smaller sets did display reduced correlation between the new pseudotime assignments for the cells subsetted and their original pseudotime values, most subsets retained a representative capability with $R > 0.9$ at 50 cells. (H) Real scRNAseq may also display zero-inflation as a result of technical dropout or transcriptional burst. By artificially adding in zero-inflation to splatter, we observed that an increase in dropout could drive a sharp decrease in pseudotime correlation. Indeed, past a certain threshold, trajectory analysis would be essentially meaningless. (I) Real scRNAseq datasets may also measure and consider gene sets of different sizes. By changing the size of a randomly selected gene set, we also observed significant changes in the distribution of pseudotime correlation values; larger datasets naturally tended to have higher noise.

interesting molecules not characterized in this context, such as the iron exporter ferroportin (SLC40A1) and zinc importer SLC39A2. Taken together, these results demonstrate that the TIPS workflow can successfully recapitulate existing knowledge of a biological context while also yielding novel candidates for further validation.

In order to put these pathway inferences relating to CD8+ T cells in context, we also further analyzed three additional scRNAseq datasets of CD8+ T cells generated using other library construction methods. From direct inspection,

we noted that of the pathways of interest we described above, four were found to be conserved across at least three of the four datasets, with particularly strong conservation of the chemokine signaling pathway we highlighted (Supplementary Figure S4A, see Supplementary Data available online at http://bib.oxfordjournals.org/). This conservation could be found despite sharp differences between datasets in terms of information recovered (Supplementary Figure S4B, see Supplementary Data available online at http://bib.oxfordjournals.org/). For instance, we could observe sharp differences in dropout rate

and background noise levels between the scSTATseq and 10X libraries (Supplementary Figure S5, see Supplementary Data available online at http://bib.oxfordjournals.org/). As such, we believe that the pathway inferences drawn through TIPS may be reproducible across multiple independent datasets.

Individual gene selection

In the results described above, we relied on picking genes with significant changes in switch-like expression as representative genes in a given pathway. However, it is well appreciated that not all genes may display this type of expression characteristic. As such, in order to develop a broader mechanism for identifying critical genes in a given pathway, we further explored using two other scoring approaches to assess gene significance. One approach is to perform a direct Pearson correlation between gene expression level for each cell with its assigned pseudotime, to capture genes with steady and monotonic expression changes. The other is to further iterate the DDRTree algorithm on a pathway-level and calculate the impact removing a single given gene would have on the strength of the pseudotime correlation (DDRTree influence). When applied to the simulated dataset described above, we found that the correlation metric was highly associated with the switchde-based results and did not help to discover more significantly associated genes (Supplementary Figure S6A, see Supplementary Data available online at http://bib.oxfordjournals.org/). However, we found that the DDRTree influence metric recovered an independent pool of genes that did not follow switch-like behavior (Supplementary Figure S6C, see Supplementary Data available online at http://bib.oxfordjournals.org/). These results were even more pronounced when applied to real data, as assessing DDRTree influence led to the identification of genes with complex expression dynamics (alternatively increasing then decreasing) (Supplementary Figure S7, see Supplementary Data available online at http://bib.oxfordjournals.org/). Since these two metrics captured distinct pools of genes that are meaningful in different context, we elected to incorporate both of these methods for gene selection to help maximize the amount of information obtainable from TIPS and assist in downstream screening.

TIPS analysis of complex progression trajectories

While differentiation trajectories may sometimes involve a simple progression from one dominant state to another, real biological trajectories are oftentimes more complex and involve multiple stable intermediate states. In order to understand if the TIPS workflow is sufficiently robust to handle these complex trajectories, we next applied it to analyze the differentiation trajectory of CD8+ tumor-infiltrating lymphocytes (TILs). TILs reside in a complex tumor-immune microenvironment, wherein different types of cellular and metabolic interactions may influence their behavior. CD8+ TILs in particular have been demonstrated to become functionally exhausted in many types of solid tumors, and prevention/reversal of exhaustion has been the focus of intensive research. To examine the processes underlying CD8+ TIL behavior, we applied the TIPS workflow to a dataset of CD8+ TILs derived from hepatocellular carcinoma (HCC) patients that were generated using the plate-based SMARTseq2 library construction.

From our initial dimension reduction via UMAP, we were able to observe a number of separate clusters that roughly corresponded to the published cellular annotations (Figure 4A).

These cell types progressed in our trajectory analysis in a somewhat irregular manner, with an undefined population of cells marking the pseudotime endpoint (Figure 4B and C). A relatively small portion of curated pathways (51/1114) were found to have significant association with this trajectory (Supplementary Table S2, see Supplementary Data available online at http://bib.oxfordjournals.org/), from which we could identify a few pathways that have been previously validated, such as receptor-tyrosine kinase signaling and TLR cascade (Figure 4D and E). From further exploration of the unannotated cluster, we found that the cluster was dominated by ribosomal signatures, explaining the inclusion of significant correlations with ribosome-related pathways in our inference (Supplementary Figure S8, see Supplementary Data available online at http://bib.oxfordjournals.org/). More interestingly, however, we also uncovered pathways that have been reported to influence CD8+ T cell behavior in other contexts, such as the ROBO receptors and ERBB2 signaling pathways [27, 28]. These pathways largely underwent substantial changes in gene expression early on in the pseudotime trajectory, although a portion of the ROBO receptor pathway shifted later on (Figure 4F). Manual inspection of the genes within these pathways demonstrated that while the expression of ROBO receptors themselves did not show strong changes with respect to pseudotime, we could observe increases in the chemokine receptor CXCR4 that has been shown to rely on ROBO cooperation (Figure 4I). Similarly, while ERBB2 itself did not show significant changes in expression, we did note a decreased expression in its downstream signal mediator KRAS at the endstage of the pseudotime trajectory (Figure 4J). Taken together, these results suggest that both of these pathways may also play significant roles in regulating CD8+ TIL behavior.

TIPS analysis of CD8+ TIL trajectory during checkpoint blockade

The recent development of ICB antibodies targeting PD-1/PD-L1 and CTLA-4 has opened a new avenue for cancer therapy. While these inhibitors are expected to have a significant impact on the behavior of TILs, the exact molecular mechanisms and processes that are altered as a result of their application are not yet fully understood. To extend our analysis of TIL behavior in HCC from above, we next analyzed a droplet-based dataset of TILs taken from patients who had undergone direct surgical resection only, or otherwise received anti-PD-L1 and anti-CTLA4 or anti-PD-1 treatment prior to resection. After subsetting out the population of CD8+ T cells from the dataset, we observed from dimension reduction that there was substantial separation between cells taken from the patients who had received differential treatment. This separation was also clear in our trajectory construction, where cells derived from untreated patients marked the end stage, while those from patients treated with anti-PD-L1 and anti-CTLA4 marked the start point.

Application of TIPS using the Reactome and KEGG knowledgebase once again yielded a small pool of (52 out of 965) pathways with significant associations (Supplementary Table S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). As a whole, these pathways were significantly different from those recovered from the previous analysis of HCC samples above. However, we were able to recover pathways such as cellular senescence and TCA cycle, which have been previously implicated to be altered as a result of ICB [29, 30] (Supplementary Figure S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). At the

same time, we also uncovered a number of pathways that have not characterized in this context, such as WNT signaling and estrogen-dependent gene expression. Whether these pathways and their downstream molecules may also be important contributors to ICB success remains an open question for future validation.

Identification of pathways involved in beta cell dysfunction in T2D

To further demonstrate the applicability of the TIPS workflow in other biological contexts, we additionally analyzed a dataset of endocrine cells derived from pancreatic islets of patients with type II diabetes (T2D) or healthy controls. UMAP clustering of the dataset cleanly stratified the seven annotated populations, with visible stratification in the beta cell cluster based on patient condition (Supplementary Figure S10A, see Supplementary Data available online at http://bib.oxfordjournals.org/). Through trajectory analysis, we were able to observe that beta cells derived from T2D patients tended to be separate from controls and occupy the end of the trajectory, a phenomenon not seen in alpha cells (Supplementary Figure S10D–I, see Supplementary Data available online at http://bib.oxfordjournals.org/). Via pathway inference, we observed contribution by pathways involved in physiological beta cell function (such as ion channel transport and IRS-related events triggered by IGF1R), as well as others associated with inflammation (such as cytokine signaling and adaptive immunity) (Supplementary Figure S10K, Supplementary Table S4, see Supplementary Data available online at http://bib.oxfordjournals.org/). Of note, direct inspection of the cytokine signaling pathway demonstrated that the genes displaying changes included downstream effectors, such as the transcription factor FOXO3 and cell death regulator CASP3, suggesting that the inference was due to the influence of cytokine stimulus on beta cells, and not direct beta cell production of cytokines (Supplementary Figure S10L, see Supplementary Data available online at http://bib.oxfordjournals.org/). These findings are consistent with previous reports and show that our TIPS analytical framework can provide pathway inferences in other settings [31].

Extension of TIPS using other trajectory inference algorithms

While the results above utilize the DDRTree algorithm for trajectory mapping, many other trajectory inference algorithms also exist, and some researchers may have preferences for alternative algorithms [32]. On a conceptual level, the TIPS workflow is not reliant on any particular trajectory inference algorithm, so long as the algorithm returns accurate pseudotime mapping results. As such, we also compared the results from application of the TIPS workflow using DDRTree with results obtainable using other trajectory inference algorithms. Using five previously benchmarked algorithms in the dynverse collection (Slingshot, Embeddr, Component1, MATCHER and Angle) for comparison, we analyzed the correlations seen between KEGG pathways and random gene lists in the scSTATseq CD8 + T cell dataset described above in Figure 3. On a general level, we observed that two approaches (Component1 and MATCHER) yielded the highest number of pathways with inferred contribution but also had the highest background signal when used to analyze randomly selected lists (Supplementary Figure S12A and B, see Supplementary Data available online at http://bib.oxfordjournals.org/). Angle, which

is intended to recover cyclical trajectories, showed no correlation with the other methods. Since the dataset itself is not expected to display cyclical behavior on a biological level (effector CD8 + T cells are not expected to have the capability to revert back into naïve CD8 + T cells), its application here is improper. Two other approaches (DDRTree and Slingshot) showed relatively similar results (Supplementary Figure S12C and D, see Supplementary Data available online at http://bib.oxfordjournals.org/), with Slingshot detecting more pathways at higher correlation values, but also having somewhat higher background noise. From these analyses, we observe that the signal-to-noise tradeoff may be an important component to consider when selecting trajectory inference algorithms for TIPS and that DDRTree use in this case appears to offer high sensitivity for pathway inference. However, the ideal algorithm in each use case may vary because of the contextual influence and other advantages/disadvantages of each algorithm (runtime, stability, accessibility, etc.). As such, we recommend users of the TIPS workflow to also consider alternative trajectory inference algorithms based on dataset fit. For reference, the runtimes using DDRTree for the datasets analyzed in this manuscript are included as Supplementary Table S5.

Discussion

While a large number of methods have been developed in recent years to help order cells along a single and/or multiple trajectories, obtaining information of biological significance of from such analysis has somewhat lagged behind [33]. In particular, although it has been demonstrated through a number of methods that trajectory analysis can recapitulate the known order of cellular maturation over the course of hematopoiesis, few studies have been able to discover novel transcription factors and/or biological processes that influence this process. In an attempt to redress this deficiency, we have presented the TIPS framework as described above to help uncover these molecular mechanisms. By relying on pseudotime trajectories as our point of comparison, we can maintain the single-cell nature of the data and thereby identify pathways that change across multiple clusters. Furthermore, we can leverage existing tools for identifying gene changes with respect to pseudotime to give an overview of the temporal order in which pathways undergo significant changes. This latter form of temporal information is an additional vantage point that may of particular use in examining interdependent pathway relationships.

Although our current pseudotime workflow is built upon the DDRTree algorithm implemented in Monocle, the TIPS workflow is not limited to this method for pseudotime assignment. Instead, the relative simplicity of our conceptual framework allows it to be readily implemented with other assignment algorithms and thus provides flexibility for further optimization through trajectory mapping algorithms with faster runtimes and improved accuracy. We hope to be able to continually update our TIPS server to incorporate these newer methods for pseudotime assignment and develop a recommendation scheme for trajectory inference algorithms based on dataset metrics in the future. Similarly, we currently rely on the switchde concept of finding genes with single, abrupt changes in expression, to define the order at which a pathway may be activating. This approach may be particularly suitable for genes with pronounced burst [34]. However, not all genes will follow this pattern of expression; some may have substantially more gradual increases in response to stimuli as a result of a higher basal level of expression, while others may display more complex expression kinetics with multiple on/off switch points.
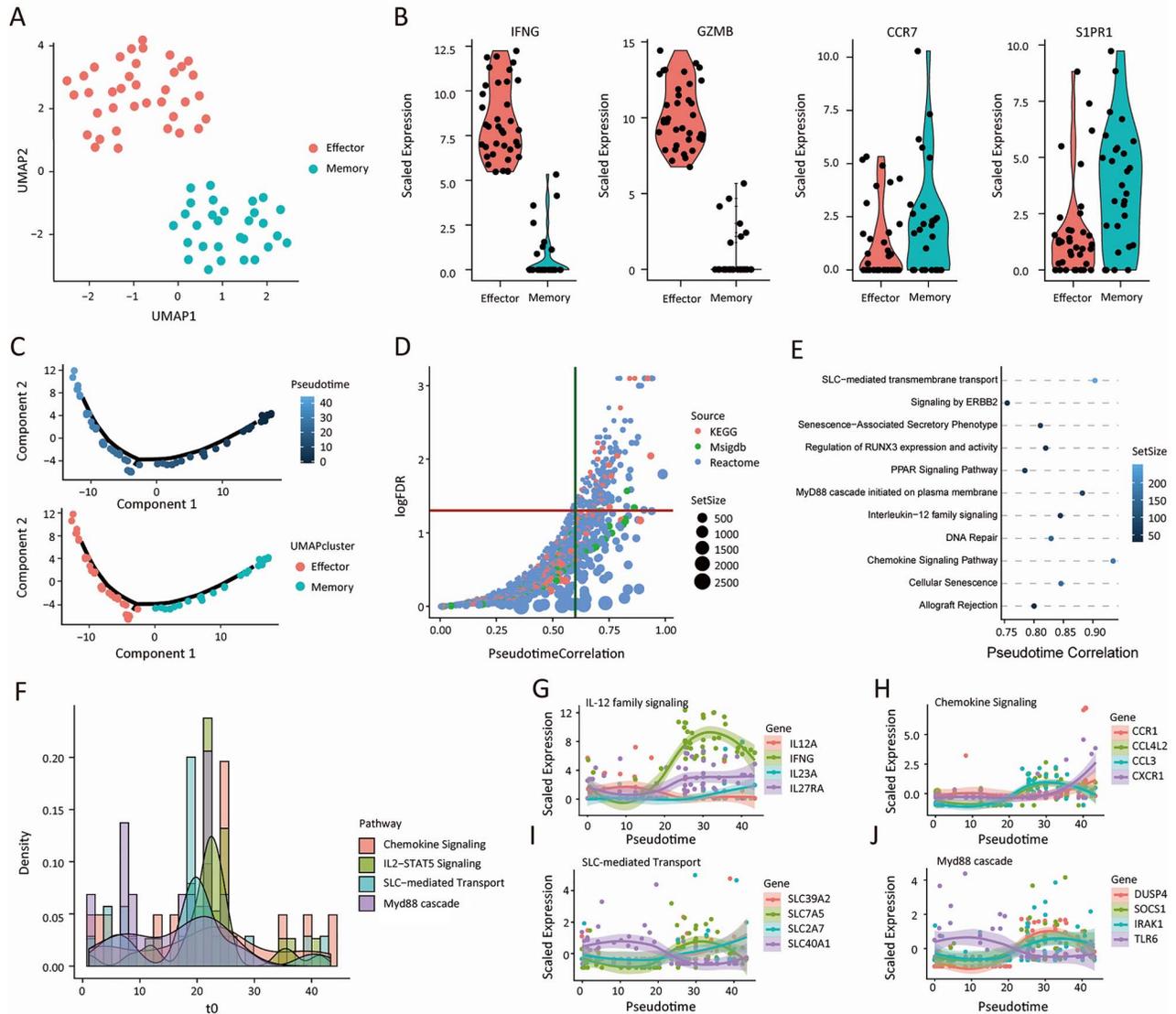
Figure 3. TIPS workflow applied to CD8+ T cells. (A) UMAP reduction and clustering of the sorted CD8+ T cells identifies two prominent clusters reflective of effector and naïve/memory like populations. (B) Violin plot of four prominent markers of T cell state demonstrate that the effector cluster features high expression of the functional molecules GZMB and IFNG, while the memory cluster shows elevated expression of CCR7 and S1PR1. (C) Trajectory mapping of these cells reveals a simple arcing path with ordered progression from memory to effector cells. (D) Volcano plot of the distribution of pseudotime correlation values and FDRs for all pathways derived from three databases considered. Notably, a number of larger pathways do display high levels of pseudotime correlation, but at a level functionally indistinguishable from randomly selected gene sets of matching size. (E) Dotplot visualization of 10 pathways of interest that had significant correlation. While TIPS includes a default option for picking the top10 pathways, users may also wish to highlight specific pathways of interest. (F) Temporal ordering of four pathways of interest using the switch points of significant genes along the trajectory. While the majority of the switch events occur at the point of change between memory and effector populations, a significant portion also occurs at earlier and later points, indicating that the changes do not simply describe DEGs between the two clusters. (G–J) Plots of the scaled expression of specific genes of interest from each pathway visualized over the course of pseudotime progression. These include genes that changed late in the trajectory, such as CXCR1 from the chemokine signaling pathway, as well as genes that change earlier on, such as IL12A from the IL-12 family signaling pathway.

Further development of gene kinetic modeling and regulatory inference algorithms may be able to identify ways to successfully capture the signal of these genes with respect to pseudotime and help refine our understanding of pathway kinetics [35]. Additional developments in systems to calibrate single-gene weight within pathway gene sets may also help to further refine the sensitivity of our kinetic inference results.

In short, our conceptual approach is not limited by the type of single-cell library construction method used; we present worked examples of TIPS analysis as performed on multiple different datasets from independent sources and generated with different workflows. These datasets vary significantly in the numbers of cells sequenced, the numbers of genes detected per cell and in their rate of technical dropout. This latter source of variation may have particularly significant ramifications on the accuracy of pseudotime alignment. A number of informatics tools have been designed to impute and correct for this variation [36]. However, as further advances in library construction lead to increasingly precise and accurate single-cell profiles, technical dropout may also be significantly ameliorated. Together with further refinement and expansion of knowledgebase data, we anticipate that the analytical framework we describe in this manuscript will only improve in accuracy and predictive power over time.
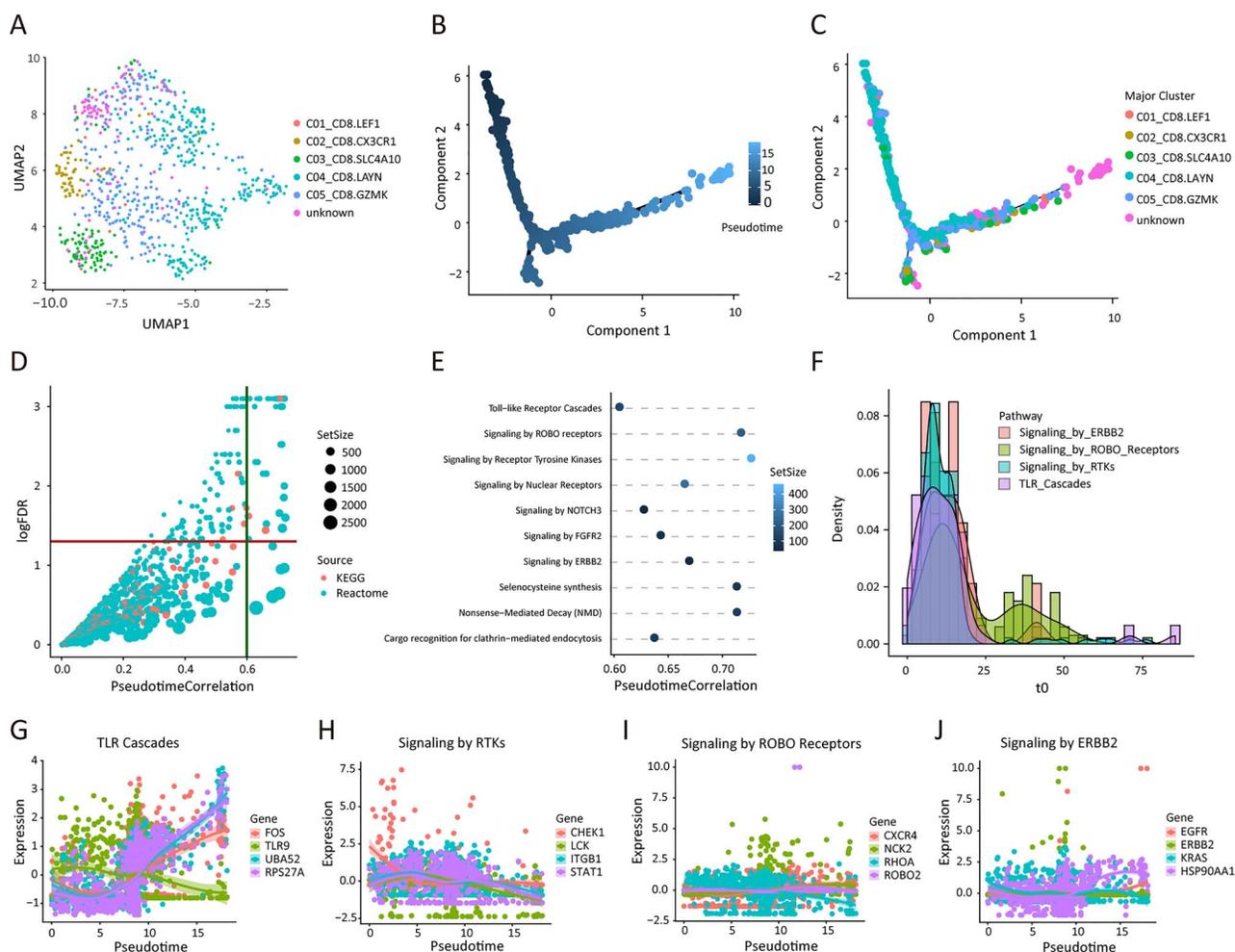
Figure 4. Application of TIPS to identify pathways associated with CD8 TIL function. (A) UMAP reduction of the subset of CD8+ TILs taken from tumor regions (annotated as TTC in original data) shows relatively clean separation between the 5 types of annotated cell populations described in the original analysis, with some overlap caused by differences in dimension reduction method and HVG selection. (B and C) Trajectory mapping of these cells shows some stratification between cell states along the pseudotime trajectory, with a relatively prominent placement of unannotated cells at the end point of the trajectory. (D) Distribution of pathways according to pseudotime correlation and FDR shows relatively few significant pathways. (E) Dotplot visualization of several significant pathways. (F) Temporal ordering of selected pathways shows that most changes occur early on along the trajectory, although the ROBO receptors pathway includes a secondary peak of changes at the middle of the trajectory. (G–J) Plots of the scaled expression of specific genes of interest from each pathway visualized over the course of pseudotime progression.

Key Points

- The TIPS framework can be used to infer which biological pathways are significantly associated with progression along a central pseudotime trajectory.
- Additional contextual information on the relative order of pathway changes, and the specific genes driving such changes, can also be identified.
- TIPS opens up new modes of information recoverable from trajectory analysis.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data Availability

All publicly available scRNAseq data analyzed in this study are available through GEO under the accessions GSE133535 (CD8+ T cell libraries constructed using other methods),

GSE98683 (SMARTseq2 sequencing of CD8+ TILs from HCC patients) and GSE125449 (10X Genomics sequencing of CD8+ TILs from HCC patients after immune checkpoint blockade). scSTATseq sequenced libraries of CD8+T cells are available through the Genome Sequence Archive (accession HRA000723).

References

1. Aibar S, González-Blas CB, Moerman T, *et al*. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**(11):1083–6. doi: 10.1038/nmeth.4463.

2. la Manno G, Soldatov R, Zeisel A, *et al*. RNA velocity of single cells. *Nature* 2018;**560**(7719):494–8. doi: 10.1038/s41586-018-0414-6.

3. Efremova M, Vento-Tormo M, Teichmann SA, *et al*. Cell-PhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* 2020;**15**(4):1484–506. doi: 10.1038/s41596-020-0292-x.

4. Lähnemann D, Köster J, Szczurek E, *et al*. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**(1):31. doi: 10.1186/s13059-020-1926-6.

5. Andrews TS, Hemberg M. False signals induced by single-cell imputation. Version 2. *F1000Res* 2018;**7**:1740. doi: 10.12688/f1000research.16613.2.

6. Baran Y, Bercovich A, Sebe-Pedros A, *et al*. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol* 2019;**20**(1):206. doi: 10.1186/s13059-019-1812-2.

7. Aibar, *et al*. AUCell: Analysis of 'gene set' activity in single-cell RNA-seq data. 2016, R/Bioconductor package.

8. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;**14**:7. doi: 10.1186/1471-2105-14-7.

9. Kanehisa M, Sato Y. KEGG mapper for inferring cellular functions from protein sequences. *Protein Sci*. Nature Biotechnol 2020;**29**(1):28–35. doi: 10.1002/pro.3711.

10. Jassal B, Matthews L, Viteri G, *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**(D1):D498–503. doi: 10.1093/nar/gkz1031.

11. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**(43):15545–50. doi: 10.1073/pnas.0506580102.

12. Butler A, Hoffman P, Smibert P, *et al*. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**(5):411–20. doi: 10.1038/nbt.4096.

13. Becht E, McInnes L, Healy J, *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**:38–44. doi: 10.1038/nbt.4314.

14. Trapnell C, Cacchiarelli D, Grimsby J, *et al*. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**(4):381–6. doi: 10.1038/nbt.2859.

15. Campbell KR, Yau C. Switchde: inference of switch-like differential expression along single-cell trajectories. *Bioinformatics* 2017;**33**(8):1241–2. doi: 10.1093/bioinformatics/btw798.

16. Wehrens R, Kruisselbrink J. Flexible self-organizing maps in kohonen 3.0. *J Stat Softw* 2018;**87**(7):1–18. doi: 10.18637/jss.v087.i07.

17. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol* 2017;**18**(1):174. doi: 10.1186/s13059-017-1305-0.

18. Zheng Z, Tang X, Qiu X, *et al*. scSTATseq: diminishing technical dropout enables core transcriptome recovery and comprehensive single-cell trajectory mapping. *Preprint: Biorxiv*. doi: 10.1101/2020.04.15.042408 April 17, 2020, preprint: not peer reviewed.

19. Patro R, Duggal G, Love MI, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**(4):417–9. doi: 10.1038/nmeth.4197.

20. Mereu E, Lafzi A, Moutinho C, *et al*. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 2020;**38**(6):747–55. doi: 10.1038/s41587-020-0469-4.

21. Zhang Q, He Y, Luo N, *et al*. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* 2019;**179**(4):829, e20–45. doi: 10.1016/j.cell.2019.10.003.

22. Ma L, Hernandez MO, Zhao Y, *et al*. Tumor cell biodiversity drives microenvironmental reprogramming in liver cancer. *Cancer Cell* 2019;**36**(4):418, e6–30. doi: 10.1016/j.ccell.2019.08.007.

23. Lawlor N, George J, Bolisetty M, *et al*. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res* 2017;**27**(2):208–22.

24. Wallin JJ, Liang L, Bakardjiev A, *et al*. Enhancement of CD8+ T cell responses by ICOS/B7h costimulation. *J Immunol* 2001;**167**(1):132–9. doi: 10.4049/jimmunol.167.1.132.

25. Hess C, Means TK, Autissier P, *et al*. IL-8 responsiveness defines a subset of CD8 T cells poised to kill. *Blood* 2004;**104**(12):3463–71. doi: 10.1182/blood-2004-03-1067.

26. Sinclair LV, Howden AJ, Brenes A, *et al*. Antigen receptor control of methionine metabolism in T cells. *Elife* 2019;**8**:e44210. doi: 10.7554/eLife.44210.

27. Prasad A, Qamri Z, Wu J, *et al*. Slit-2/Robo-1 modulates the CXCL12/CXCR4-induced chemotaxis of T cells. *J Leukoc Biol* 2007;**82**(3):465–76. doi: 10.1189/jlb.1106678.

28. Lozano T, Chocarro S, Martin C, *et al*. Genetic modification of CD8(+) T cells to express EGFR: potential application for adoptive T cell therapies. *Front Immunol* 2019;**10**:2990. doi: 10.3389/fimmu.2019.02990.

29. Im SJ, Hashimoto M, Gerner MY, *et al*. Defining CD8+ T cells that provide the proliferative burst after PD-1 therapy. *Nature* 2016;**537**(7620):417–21. doi: 10.1038/nature19330.

30. LaFleur MW, Nguyen TH, Coxe MA, *et al*. PTPN2 regulates the generation of exhausted CD8(+) T cell subpopulations and restrains tumor immunity. *Nat Immunol* 2019;**20**(10):1335–47. doi: 10.1038/s41590-019-0480-4.

31. Eguchi K, Nagai R. Islet inflammation in type 2 diabetes and physiology. *J Clin Invest* 2017;**127**(1):14–23. doi: 10.1172/JCI88877.

32. Wolf AF, Hamey FK, Plass M, *et al*. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**(1):59. doi: 10.1186/s13059-019-1663-x.

33. Saelens W, Cannoodt R, Todorov H, *et al*. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**(5):547–54. doi: 10.1038/s41587-019-0071-9.

34. Larsson AJM, Johnsson P, Hagemann-Jensen M, *et al*. Genomic encoding of transcriptional burst kinetics. *Nature* 2019;**565**(7738):251–4. doi: 10.1038/s41586-018-0836-1.

35. Pratapa A, Jalihal AP, Law JN, *et al*. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods* 2020;**17**(2):147–54. doi: 10.1038/s41592-019-0690-6.

36. Lan T, Hutvagner G, Lan Q *et al*. Sequencing dropout-and-batch effect normalization for single-cell mRNA profiles: a survey and comparative analysis. *Brief Bioinform* 2020;bbaa248. doi: 10.1093/bib/bbaa248.