scRegulocity: Detection of local RNA velocity patterns in embeddings of single cell RNA-Seq data

Akdes Serin Harmanci[1]+, Arif O Harmanci[2]+*, Xiaobo Zhou[2], Benjamin Deneen[1,3,4], Ganesh Rao[1], Tiemo Klisch[5,6], Akash Patel[1,5]*

Affiliations:

1 Department of Neurosurgery, Baylor College of Medicine, Houston, TX 77030
2 Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, 77030
3 Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, TX 77030
4 Program in Developmental Biology, Baylor College of Medicine, Houston, TX 77030
5 Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX 77030
6 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030
+ These authors contributed equally
* Corresponding authors

Abstract

Single cell RNA-sequencing has revolutionized transcriptome analysis. ScRNA-seq provides a massive resource for studying biological phenomena at single cell level. One of the most important applications of scRNA-seq is the inference of dynamic cell states through modeling of transcriptional dynamics. Understanding the full transcriptional dynamics using the concept named RNA Velocity enables us to identify cell states, regimes of regulatory changes in cell states, and putative drivers within these states. We present scRegulocity that integrates RNA-velocity estimates with locality information from cell embedding coordinates. scRegulocity focuses on velocity switching patterns, local patterns where velocity of nearby cells change abruptly. These different transcriptional dynamics patterns can be indicative of transitioning cell states. scRegulocity annotates these patterns with genes and enriched pathways and also analyzes and visualizes the velocity switching patterns at the regulatory network level. scRegulocity also combines velocity estimation, pattern detection and visualization steps.

Introduction

Single-cell RNA Sequencing has enabled us to study heterogeneous cell populations with single cell resolution. Technologies such as 10X Genomics Chromium[1], inDrop[2], SMART-seq2[3], Drop-Seq[4] are used to sequence transcripts from thousands of cells that are isolated from a sample of interest. Current data analysis pipelines analyzing scRNA-Seq data reveals a static snapshot of cellular states. Standard scRNA-seq pipelines focus on quality control[5–9], cell filtering [5–9], dimensionality reduction [5–9], integration[10–17], differential expression[18–20], and clustering of the cells[5,8,9], and assignment of cell types from the samples[21–25]. These are very important steps to organize and assess the quality of the single cell datasets and provide an initial analysis of the data. As single cell datasets tend to be very large with possibly hundreds of thousands of cells, these initial analyses provide important insight into the biological states of the cells and the studied conditions. As such, the single cell datasets contain massive amount of information that

should be extracted with computational and statistical methods. One of the main challenges is that datasets are being generated at a rate that is much faster than the computational methods can analyze.

Although scRNA-seq is generated from a single time point, it can be used to estimate the transcriptional dynamics of transcriptional states of cells to study, for example, developmental, disease, and immunological processes that exhibit large dynamic changes in cells[26–29]. Understanding transcriptional states enables us to define cell types and cell-type defining markers more coherently. Additionally, it allows us to infer the heterogeneity of tumor samples at the single cell level.

One way to infer transcriptional dynamics is through trajectory analysis. The main hypothesis for these analyses is that the sample comprise heterogeneous sets of cells from a continuum of dynamic states. These states can represent dynamic processes such as differentiation and cell cycle[30–35]. These states can be modeled by numerous "trajectories" where the dynamic states are connected on a trajectory of states (e.g., Markov processes). Methods that perform trajectory analysis assume that the continuum of cellular states are sufficiently observable in the single cell RNA-seq sample. The idea is to build parsimonious trajectories that explain the changes in the cell types. The trajectory analysis is often coupled with pseudotime analysis[36,37] to assign relative time units to the dynamic trajectory of the cells. This way, the cells in each of the trajectories can be aligned properly.

There are also methods that extract dynamicity information from all the cells at the same time by estimating the derivative of gene expression levels. This is performed by concept named RNA velocity, that can reliably estimate the relative time derivative of the gene expression state. RNA velocity enables us to study cellular transcription kinetics using the ratio of spliced and unspliced read counts of each gene across RNA-Seq data [28,29,38,39]. The underlying assumption in this model is that genes are initially transcribed in an unspliced manner and then spliced, such that observed intronic reads can be interpreted as corresponding to nascently transcribed mRNAs. Transcriptional upregulation of a gene will result in a transient excess of nascent (unspliced) transcripts compared with processed (spliced) transcripts, whereas transcriptional downregulation results in a relative depletion of nascent (unspliced) transcripts.

Unlike trajectory analysis, RNA-velocity does not require the cellular composition to be diverse enough since each cell is processed by itself and velocity can, in principle, be estimated in each cell independently. Although several methods have utilized RNA-velocity for building trajectories and estimating cellular dynamics at a sample-wide (or global) level, there is still much information to be extracted from "local patterns", i.e. the interactions of subsets of cells have with each other. The "local patterns" can be more concretely described by considering embeddings of cells in lower dimensions where the nearby cells in the embeddings are more similar to each other in terms of transcriptional states. One example of these is tSNE and UMAP based embeddings that are used extensively for visualizing scRNA-seq datasets. The local patterns in the embeddings can provide an incredible amount of biological insight. The relationship between velocity and the cellular dynamicity at the level of localities in the

embeddings is not well-studied. We hypothesize that there is a need to develop new computational methods for detecting, summarizing, and visualizing the biological insights of dynamicity of cellular states in connection to the embeddings.

In this study, we present the scRegulocity algorithm, for measuring the dynamics of gene expression in large numbers of single cells using RNA velocity. Specifically, scRegulocity detects genes with velocity switches whereby the gene exhibits a strong velocity difference among cells that are nearby in the coordinates of embeddings. The local velocity switching patterns are very frequently observed in manual inspection of the estimated velocity distributions on the cells. We believe that the genes with velocity switches potentially represent drivers of dynamic processes such as cellular differentiation/development and disease progression, and these can be instrumental to delineate the drivers of dynamicity of these processes especially in tumors and cancers. In particular, these genes exhibit transcriptional dynamics which are detected by integrating RNA velocity and expression (to build the embeddings) rather than using expression signatures alone. ScRegulocity also reconstructs gene regulatory networks in transitioning cell states using RNA velocity. Our analyses on different single cell RNA-Seq datasets show that scRegulocity can recover driver TFs and transcriptional programmes in transitoning cell states which can not be easily inferred from whole transcriptome data. We believe that scRegulocity will facilitate the study of gene regulation in diverse biological systems.

Compared to other methods, scRegulocity stands out as a “local dynamicity inference” tool, whereas the majority of the other tools aim at detecting and describing patterns at a sample-wide level (or globally). scRegulocity takes standard files as inputs, is flexible and can be integrated into scRNA-seq analysis pipelines.

Results

scRegulocity Algorithm

Figure 1 illustrates the scRegulocity algorithm. The input is the aligned RNA-seq reads (e.g., SAM/BAM file) and the list of cell ids that will be analyzed. scRegulocity provides an integrated and complete pipeline starting from mapped reads and uses a spatial signal processing approach to detect the velocity switching patterns on embeddings. A velocity switching pattern is defined by an abrupt coordinated increase (or decrease) in velocity between two groups of cells that are close in the embeddings. Thus, it is vital for the embedding of cells into lower dimensions to provide useful biological information for nearby cells that are close to each other. For most of the embeddings that are widely used (such as tSNE, UMAP, and PCA) closeness generally implies biological similarity and therefore should be meaningfully usable in the context of velocity-switching analysis. In this study, we focus on tSNE and UMAP-based dimensionality reduction using the gene expression counts, i.e., the embeddings represent similarities in the global transcriptomic profiles. The velocity switching patterns are expected to identify the cells that are similar in transcriptional state but harbor opposite dynamic changes in expressional states that potentially stem from the regulatory state of the cells. One of the motivations for developing scRegulocity is that the velocity switching patterns are frequently observed in manual inspection of the expression velocities after they are mapped on the embedding coordinates.

In order to comprehensively characterize the velocity-switching patterns at the regulatory level, we mapped known TF-target interactions onto the genes that exhibit significant velocity switching-patterns. Then scRegulocity classifies the concordance or discordance of velocity switches with regulatory relations between the TF and target at the velocity-level and/or expression-level. Finally, the identified regulatory switches and regulatory information are visualized on the embeddings. scRegulocity can generate the popular embeddings of the cells after quantifying the expression levels on each cell. However, the user can skip this step if he/she has already generated the embedding themselves. We describe the other steps of scReguloCity workflow in Methods section.

Accuracy of velocity estimates using sci-fate data

We first applied scRegulocity algorithm on cortisol response dataset generated from a method named sci-fate[40]. Sci-fate method is a combined single-cell combinatorial indexing and mRNA labelling to profile the 'older' and 'newer' transcripts based on their splicing status in single cell resolution. In this study, researchers identified regulatory elements and transcriptional drivers in cortisol response using the newly synthesized expression values. The newly synthesized expression values in scifate-study is the ground truth for the RNA Velocity values. Therefore we validated our scRegulocity algorithm using the cortisol response dataset and detected similar transcriptional programs reported in the scifate study.

In order to show the similarity of velocity and newly synthesized expression, we first calculated the correlation of velocity with newly synthesized and whole-transcriptome data for each gene separately. Figure 2A shows the distribution of correlation for each gene. We have detected a higher mean correlation with velocity and newly synthesized data compared to whole-transcriptome data (Figure 2A). We also checked the correlation of velocity of TF with TF target genes reported in sci-fate study (Figure 2B). We observed a higher correlation between velocity of TF with its target gene velocity values. Thus, we can infer the TF target regulatory network more accurately using RNA velocity values.

We next identified genes that have uniform direction of the velocity switch vectors in a subset of cells in order to define driver genes in cell state transitions. The velocity of cell cycle TFs such as POLR2A, NF1, BRCA1 and GR response TFs such as TEAD1 were highly correlated with the levels of velocity, more so than overall target gene mRNAs (Figure 2C).

scRNA-seq of Chromaffin differentiation

We next applied our scRegulocity algorithm on a Chromaffin differentiation dataset studied in velocyto paper[41]. Researchers detected a movement of the differentiating cells towards a chromaffin fate using RNA Velocity. We first detected genes with velocity switches among different cell states and types. ScRegulocity identified Serpine2 having a significant velocity switching pattern among SCP (schwann cell precursor) cells and Differentiation cells (Figure 3A). We next clustered genes using velocity values and identified different velocity switching patterns (Figure 3B). Then we performed enrichment analysis on the genes within each cluster. We next sought the TFs that drive the progression of chromaffin fate differentiation, and inferred TF target regulatory network using RNA velocity values (Figure 3C). Chromaffin differentiation related TFs, such as Gata3, Tcf7l2, Sox6 and Tcf4, were identified using

scRegulocity and the TF velocity values of these TFs were highly correlated with the mean TF target velocity values (Figure 3D).

Single-cell RNASeq Meningioma

We also applied our algorithm on our scRNA-Seq meningioma dataset. In total, after filtering low quality cells we have n=12244 cells from n=2 NF2 mutant recurrent meningioma tumors. We first clustered our scRNA-Seq data using the Louvain community detection algorithm. This generated a total of n=16 clusters. We next annotated the clusters with cell types using singleR algorithm and well-established cell type markers. We identified monocyte, macrophage, T-cell and tumor cells (Figure 4A). We next identified large scale CNV events using CaSpER[42] to elucidate the effect of CNVs on velocity (Supplementary Figure 1). We noticed the tumor cluster 0 and 8 not harboring chr11p and chr18q deletion. We believe that these cells represent less aggressive cell clones within the tumor compared to other tumor cell clusters. We supported our hypothesis by calculating an aggressiveness score for each cell using gene signatures of aggressive and non-aggressive tumors identified from our previously published bulk RNA-Seq expression data[43]. We have observed that cluster 0 and cluster 8 got a higher score for non-aggressive meningioma tumors. We next inferred RNA velocity in our scRNA-Seq meningioma data and projected the velocities on to our previously defined UMAP embeddings. We observed a similar finding that non-aggressive meningioma cells are moving towards aggressive meningioma cells in velocity based trajectory analysis (Figure 4B).

We next applied our scRegulocity algorithm on our single cell meningioma data. We observed that hypoxia related TFs such as *DDIT3 and NR3C1* have repressed transcriptional dynamics in less aggressive tumor cells. Similarly, *TCF7L1* which is a mediator of the Wnt signaling pathway[44], and *CDK2AP1* which epigenetically regulates embryonic stem cell differentiation[45], have increased transcriptional dynamics in more aggressive tumor cells (Figure 4C).

Methods

Intron/Exon Read Quantification. The velocity estimation starts by quantifying the intron/exon read counts. The basic idea is that genes that exhibit increase (decrease) in expression will harbor more (less) reads on the introns compared to the baseline exonic reads. This basic idea is used to estimate and assign expression velocity estimates to each gene. scRegulocity contains a module that counts intronic (unspliced) and exonic (spliced) read counts for each gene in each sample. and normalize the counts using total number of reads in each sample. scRegulocity has a specific module to perform read quantifications in an integrated manner so that there is no dependence on the other packages. Specifically, scRegulocity makes use of the "CB:Z" tags in the reads to first assign each read to a cell then identified whether the read belong to an intron, exon, or an intron-exon junction, i.e., unspliced reads. scRegulocity keeps track of 3 different counters ($R_{g,c}^{(exon)}, R_{g,c}^{(intron)}, R_{g,c}^{(intrex)}$, corresponding to exonic, intronic and boundary read counts for cell at index c, and the gene at index g) for each gene g and concurrently keeps track of these counts while quantification is being performed. After quantification is finished, the count matrix (genes in the rows, cells in the columns) is saved in a tab-delimited file. The

C++ code for quantification, can be downloaded from GitHub at https://github.com/harmancilab/IntrExtract/. Our method takes a bam file as an input and outputs 3 matrices where the rows are the genes and the columns are exons, introns or ambigious reads.

Velocity Estimation. scRegulocity includes a flexible and integrated velocity estimation module that can be parametrized by the user. The velocity estimation takes the intron/exon read counts matrix as input. The number of reads mapping to the introns and intron-exon boundaries are generally 1-2 orders of magnitude smaller than that of reads mapping to the exons. This is expected since exonic reads dominate the transcripts that are sequenced in RNA-seq protocols. For velocity estimation, it is necessary to obtain a robust estimate of the ratio of reads that are mapping on introns (and intron/exon boundaries) and the exonic reads, which is proportional to the expression velocity. To provide an estimate of this ratio, scRegulocity performs a linear regression between the unspliced ($R_{g,c}^{(intron)}$) and spliced read counts of all genes ($R_{g,c}^{(exon)}$), as it is currently a well-established approach to estimate velocity [28,29]:

$$R_{g,c}^{(intron)} = a \times R_{g,c}^{(exon)} + \epsilon + \nu_{g,c}$$

where intronic read counts are modeled as an ordinary linear model of the exonic read counts, a indicates the slope, ϵ represents a random noise term to include technical noise, and $\nu(g)$ represents the velocity of the expression. From this model, the general linear trend between unspliced and spliced reads quantifies a gene-specific spliced/unspliced read counts. This effect represents mostly a technical component (see above) whereby the highly expressed genes will contain more reads on the introns and is removed by subtracting the linear component from the spliced/unspliced ratios of the genes. The residual intronic (unspliced) read counts are used as the final velocity estimates for all genes. To make the estimate more robust, the linear model uses extremes of the spliced/unspliced ratios, specifically the upper and lower quantiles, which is set to $\%q_v = 0.05$ by default, an approach similar to the velocyto workflow [46]. This parameter can be changed to make the linear trend removal more stringent or more relaxed. To test for other factors that may bias velocity estimates, we have tested the model by including covariates such as read-mappability and GC content. We observed that these covariates do not significantly improve velocity estimates and therefore are by default not explicitly included in the velocity estimation module of scRegulocity. The final velocity estimation step will yield a matrix of RNA velocity estimates where each row represents a gene and column represents cells.

Building the Cell-Cell Gradient Graph. scRegulocity identifies the velocity switching patterns using a graph-based approach. The target is to identify two sets of cells that are neighboring in the embedding such that there is a coordinated switch in the expression velocities of all cells from one set of cells to the other set of cells. In other words, we would like to identify a strong coordinated gradient between two sets of neighboring cells such that the velocities are switched between the two sets of cells. First, the embedding coordinates of the cells are analyzed and a pairwise cell-cell distance matrix is generated. Given a K-dimensional embedding, this can be simply computed by:

$$\Delta_{c,d} = \sqrt{\left(\sum_{1 \le k \le K} \left|E_{c,k} - E_{d,k}\right|^2\right)}$$

Where $E_{c,k}$ denotes the embedding coordinates of cell c at coordinate k. Based on the cell-cell distance matrix, scRegulocity uses a neighborhood parameter σ_v that denotes the largest radius at which cells are deemed as neighbors of each other. For each neighborhood, scRegulocity forms a graph where the nodes are placed on cells and edges are placed between cells that are in each other's σ_v-neighborhoods. For each edge, we assign a weight based on the absolute value of velocity difference between the cells that are connected by the edge. Given the velocity estimates for the gene g,

$$e_{c,d}^{(g)} = \left|v_{g,c} - v_{g,d}\right|; \forall c, d;\ \Delta_{c,d} < \sigma_v$$

where $e_{c,d}^{(g)}$ represents the weight of the edge that connects cells c and d, which are in the σ_v-neighborhood of each other. The edges are also assigned directions based on the sign of difference between the velocities of the cells. In this representation, the edges represent discrete units of velocity gradients that will be used to detect concordant velocity-switches. However, we observed that most of the edges do not provide useful information as they represent random and weak gradient vectors between cells. In addition, processing of the cell-cell network with all the edges increases computational cost. To overcome this, the edges are pruned with respect to the weight threshold, e_{min}, weights so that the weak gradients are excluded from analysis, i.e., for a gene g, the edges $e_{c,d}^{(g)} > e_{min}$ are retained from edge filtering. Currently, $e_{min} = 2$ is used by default as a stringent weight threshold that also provides enough power to identify velocity switching patterns. After the edges are filtered, the weakly connected components of the graph (all cells are connected to each other without regard to the direction of edges) are identified using breadth-depth first search algorithm[47]. We refer to these components as candidate velocity switching subnetworks because they contain the sets of cells where the velocity switching events take place. For a gene g, Each candidate is defined by the connected cell-cell edge subnetwork with the filtered edges:

$$N_i^{(g)} = \left\{\left(c, d, e_{c,d}^{(g)}\right)\right\}, e_{c,d}^{(g)} > e_{min}, \text{nodes in } N_i^{(g)} \text{ are weakly connected}$$

where $N_i^{(g)}$ denotes i^{th} subnetwork for gene g.

Detection of Velocity-Switching Patterns. The cell-cell gradient information in each network is expected to correspond to one velocity switching pattern. In order to detect abrupt changes in velocity, the candidate subnetworks are tested to identify whether there is a significant velocity switching pattern in them. The first test checks for concordance of the gradient in the subnetwork. For this, scRegulocity computes the directions of the gradients defined by each edge. The orientation of the gradient vectors are assigned using the coordinates of the cells that they are connecting. Given the $N_i^{(g)}$, the list of all edges in the i^{th} cell-cell subnetwork, the direction and weights of all the edges in the subnetwork are extracted:

$$\Omega\left(N_i^{(g)}\right) = \left\{\left(e_{c,d}^{(g)}, \arctan\left(\frac{E_{c,2} - E_{d,2}}{E_{c,1} - E_{d,1}}\right)\right)\right\}$$

where $\Omega\left(N_i^{(g)}\right)$ denotes pairs of edge weights and the edge orientation angle, which is computed using inverse tangent (i.e., $\arctan$) of the height/width ratio of the rectangle formed

on the embedding coordinates of the cells at the ends of edge. scRegulocity also uses the orientation and weights of the edge to compute the aggregated gradient vector, which is used in visualization:

$$V\left(N_{\mathrm{i}}^{(g)}\right) = \sum_{\left(c,d,e_{c,d}^{(g)}\right)} e_{c,d}^{(g)} \cdot (\sin(\theta), \cos(\theta)), \theta = \arctan\left(\frac{E_{c,2} - E_{d,2}}{E_{c,1} - E_{d,1}}\right)$$

This vector is a 2-element vector of embedding coordinates that is vector summation of the unit vectors along each gradient vector after the unit vectors are multiplied by the weight of the corresponding edge. $\Omega\left(N_i^{(g)}\right)$ contains the direction and strength information of the cell-cell velocity gradients. scRegulocity uses this information to statistically test whether there is an enrichment of high weighted cell-cell gradients along the same orientation. For this, Rayleigh test, which performs a statistical test of whether the sample of orientations are different from a random distribution that are sampled uniformly over the unit circle. Rayleigh test also considers the weights naturally to weigh each orientation so that the gradient vectors with higher weights contribute more on the test statistic. For a subnetwork $N_i^{(g)}$, we input $\Omega\left(N_i^{(g)}\right)$ into the Rayleigh test to assess the significance of whether the weighted edges are distributed uniformly. The subnetworks are filtered with respect to the p-value threshold. As a further test of the significance of a gradient pattern on the subnetwork, scRegulocity computes the spatial correlation using moranI test on RNA velocity gradient vectors between the cells near the velocity switch. This test takes as input the embedding coordinates of the cells and the velocity values on each cell, i.e., $v_{g,c}$. The significance of the pattern on the embedding are computed with respect to a null model where the velocities are randomly distributed.

scRegulocity uses the p-value of the moranI test to filter out subnetworks that do not exhibit significant spatial patterns. After this step, the subnetworks are scored and filtered and we have the final set of velocity switching patterns at the cells that are contained in the filtered subnetworks. For each gene, a number of subnetworks are identified.

Clustering of genes with respect to velocity switch patterns. We cluster genes based on velocity switch patterns. For this scRegulocity build a velocity direction matrix for each subnetwork for each gene. The matrix entries contain the positive (+1) and negative (-1) values indicating the direction of velocity in all cells. The matrix contains genes in the rows and cells in the columns. The genes are clustered with respect to similarity of the (discretized) velocity direction values using Euclidean distance matrix with partitioning around medoids clustering method[48]. We hypothesize that the genes that share the velocity switch pattern on same cells exhibit similar abrupt coordinated changes in the gene regulatory processes. To uncover this, scRegulocity performs pathway enrichment analysis on the set of genes within each cluster using enrichR R package[49].

Transcription Factor (TF)-Target Velocity Regulation. For each TF gene with at least one significant velocity switching pattern, scRegulocity evaluates the targets (via motif and regulatory-network databases [50]. Next, the expression levels and velocities of the targets are checked for correlation with the regulator's expression levels and velocities. These are visualized in a network view at the velocity and expression level.

Visualization. scRegulocity contains visualization modules that automatically take the outputs of the algorithm as input and generate visualizations of (1) the significant velocity switching

patterns with a directional distribution of gradients on subnetworks, (2) gene-clusterings, and (3) TF-target networks are output and visualized by scRegulocity so that the results can be easily interpreted and assessed in terms of biological significance with respect to the tested hypotheses.

Discussion

We present an algorithm, scRegulocity, for identification and visualization of driver genes and regulatory networks within transient cell states and types. We demonstrated that scRegulocity can deconvolute the transcriptional drivers using RNA Velocity with our graph based algorithm. We present several examples where scRegulocity effectively complements the existing set of RNA Velocity analysis tools and gives insight into the understanding of cell-state transitions in diverse systems. ScRegulocity can extend the utility of RNA-seq datasets beyond just transcriptional profiling.

In conclusion, scRegulocity is a method that generates RNA Velocity from single cell RNA-seq data and infers driver transcription factors and transcriptional modules to guide the discovery and understanding of the cellular states. Our results show that scRegulocity can more accurately recover dynamic transcription factor (TF) modules compared to whole transcriptome single cell expression RNA data.

Acknowledgement

This study is partially funded by the NIH grant R01CA241930.

Figure Legends

Figure 1. Overview of scRegulocity algorithm

Figure 2. A) Correlation of velocity with newly synthesized and whole-transcriptome data for each gene separately **B)** Correlation of velocity of TF with TF target genes reported in sci-fate study **C)** RNA Velocity of cell cycle TFs such as *POLR2A*, *NF1*, *BRCA1* and GR response TFs such as *TEAD1* were highly correlated with the levels of velocity.

Figure 3. **A)** ScRegulocity identified Serpine2 having a significant velocity switching pattern among SCP (schwann cell precursor) cells and Differentiation cells **B)** velocity switching patterns identified by ScRegulocity **C)** identified TF target regulatory network using RNA velocity values **D)** Chromaffin differentiation related TFs, such as *Gata3*, *Tcf7l2*, *Sox6* and *Tcf4*, were identified using scRegulocity and the TF velocity values of these TFs were highly correlated with the mean TF target velocity values.

Figure 4. A) scRNA-Seq meningioma clustering and annotation **B)** trajectory of scRNA-Seq meningioma data using RNA Velocity and aggressiveness score for each cell using gene signatures of aggressive and non-aggressive tumors **C)** *TCF7L1* which is a mediator of the Wnt signaling pathway, and *CDK2AP1* which epigenetically regulates embryonic stem cell differentiation, have increased transcriptional dynamics in more aggressive tumor cells.

References

1. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 1–12 (2017).
2. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
3. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, (2013).
4. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
5. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
6. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. *Bioinformatics* **35**, 4472–4473 (2019).
7. Melsted, P. *et al.* Modular and efficient pre-processing of single-cell RNA-seq. *bioRxiv* 673285 (2019) doi:10.1101/673285.
8. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
9. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).
10. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289–1296 (2019).
11. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36**, 421–427 (2018).
12. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018).
13. Leek, J. T. Svaseq: Removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Research* **42**, e161 (2014).
14. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
15. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).
16. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology* **37**, 685–691 (2019).
17. Barkas, N. *et al.* Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nature Methods* **16**, 695–698 (2019).
18. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278 (2015).
19. Kharchenko, P. v., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740–742 (2014).
20. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

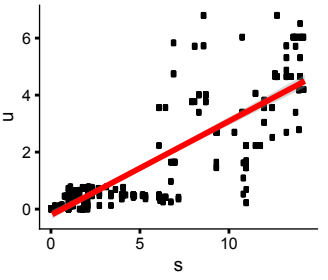
21. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**, 163–172 (2019).
22. Miao, Z. *et al.* Putative cell type discovery from single-cell gene expression data. *Nature Methods* **17**, 621–628 (2020).
23. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology* **20**, 1–17 (2019).
24. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic acids research* **47**, e95 (2019).
25. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods* **15**, 359–362 (2018).
26. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine* **26**, 1070–1076 (2020).
27. Couturier, C. P. *et al.* Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nature Communications* **11**, 1–19 (2020).
28. la Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
29. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38**, 1408–1414 (2020).
30. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**, 547–554 (2019).
31. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* **20**, 1–9 (2019).
32. Cannoodt, R., Saelens, W. & Saeys, Y. Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology* vol. 46 2496–2506 (2016).
33. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381–386 (2014).
34. Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology* **34**, 637–645 (2016).
35. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13**, 845–848 (2016).
36. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**, 381–386 (2014).
37. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nature Methods* **14**, 309–315 (2017).
38. Svensson, V. & Pachter, L. RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq. *Molecular Cell* vol. 72 7–9 (2018).
39. Zeisel, A. *et al.* Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology* vol. 7 529 (2011).
40. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nature Biotechnology* **38**, 980–988 (2020).
41. la Manno, G. *et al.* RNA velocity of single cells. *Nature* vol. 560 494–498 (2018).

42. Serin Harmanci, A., Harmanci, A. O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nature Communications* (2020) doi:10.1038/s41467-019-13779-x.
43. Patel, A. J. *et al.* Molecular profiling predicts meningioma recurrence and reveals loss of DREAM complex repression in aggressive tumors. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 21715–21726 (2019).
44. Shy, B. R. *et al.* Regulation of Tcf7l1 DNA Binding and Protein Stability as Principal Mechanisms of Wnt/β-Catenin Signaling. *Cell Reports* **4**, 1–9 (2013).
45. Deshpande, A. M. *et al.* Cdk2ap1 Is Required for Epigenetic Silencing of Oct4 during Murine Embryonic Stem Cell Differentiation S. (2009) doi:10.1074/jbc.C800158200.
46. la Manno, G. *et al.* RNA velocity of single cells. *Nature* vol. 560 494–498 (2018).
47. Csárdi, G. & Nepusz, T. *The igraph software package for complex network research*.
48. Hennig, C. & Liao, T. F. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **62**, 309–369 (2013).
49. Kuleshov, M. v. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016).
50. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods* **14**, 1083–1086 (2017).

INPUT: Aligned Reads (BAM files)

Single-cell RNA-Seq

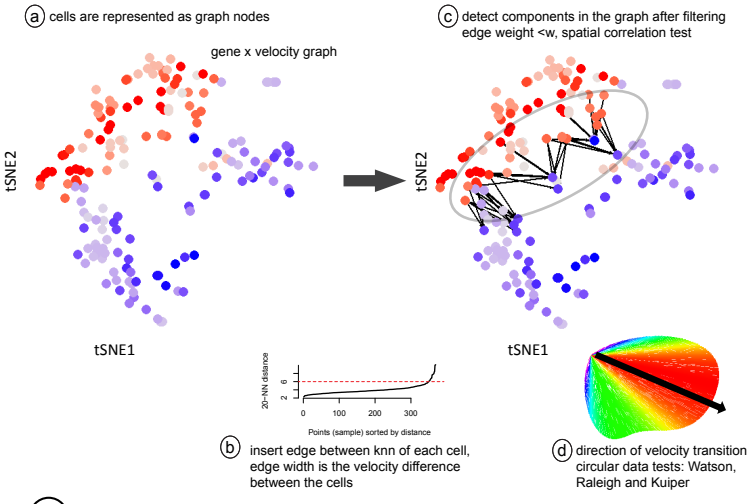
GENE LEVEL VELOCITY ESTIMATES



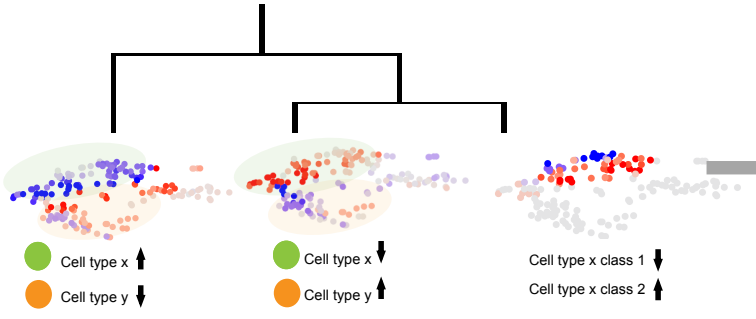
1. Extract normalized intron and exon read counts
2. Linear fit on unspliced and spliced normalized counts for each gene

A. IDENTIFY TRANSITIONING CELL STATES

1. Identify cells in the border of velocity switch

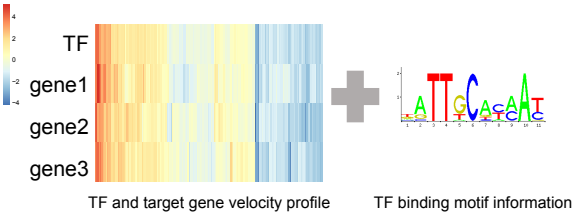


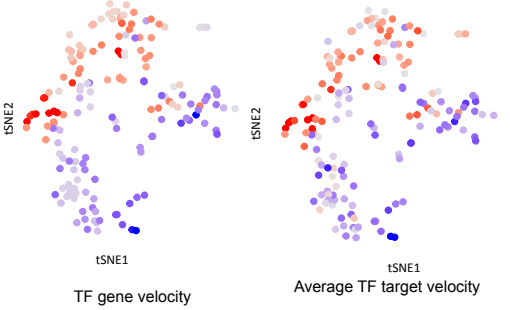
2. Cluster genes based on velocity switch patterns



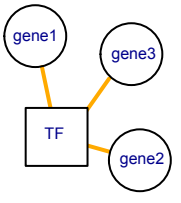
B. IDENTIFY TF GENE REGULATORY NETWORK

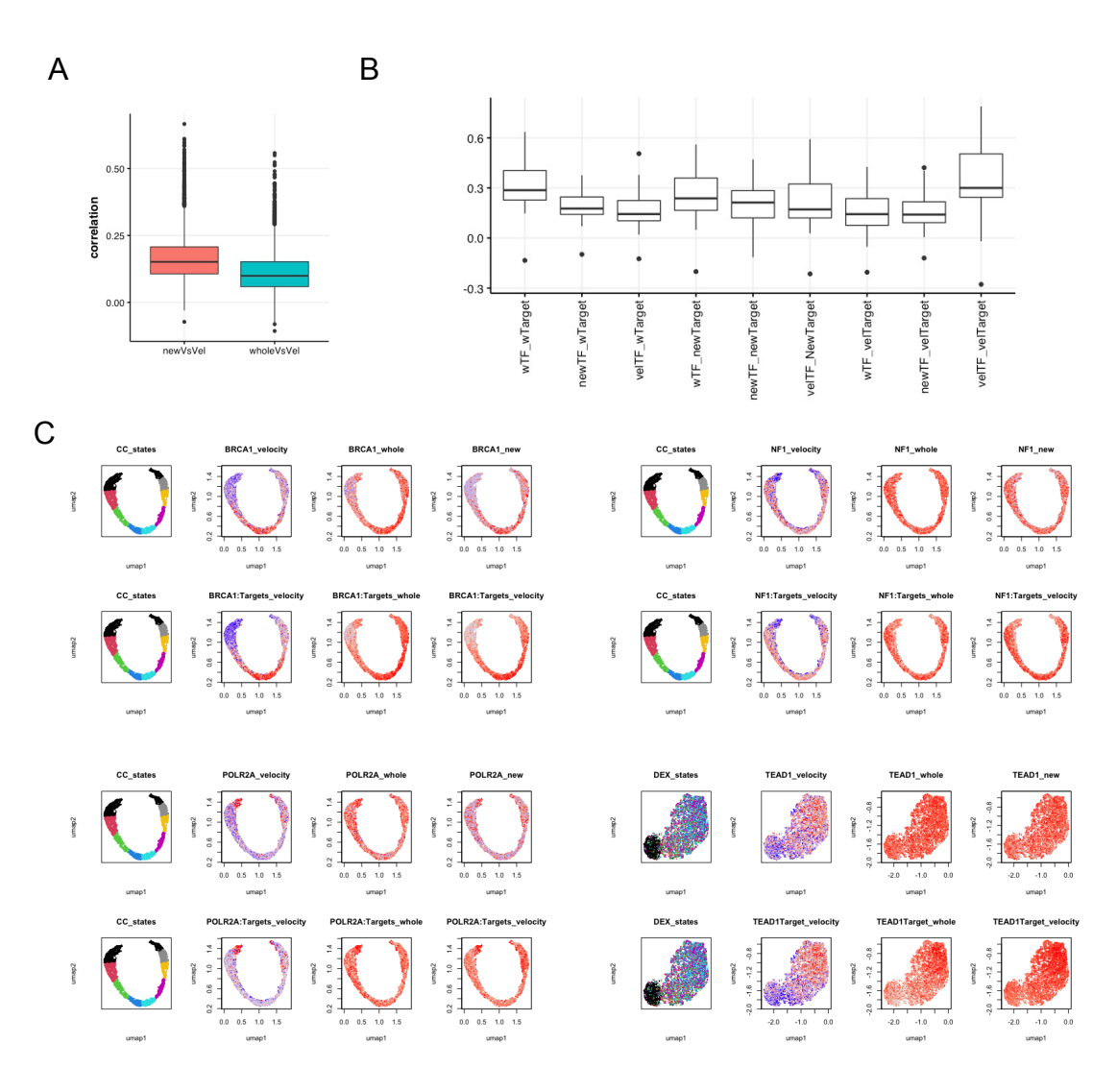
3. Identify TFs and genes in each velocity switch patterns and construct TF co-velocity modules

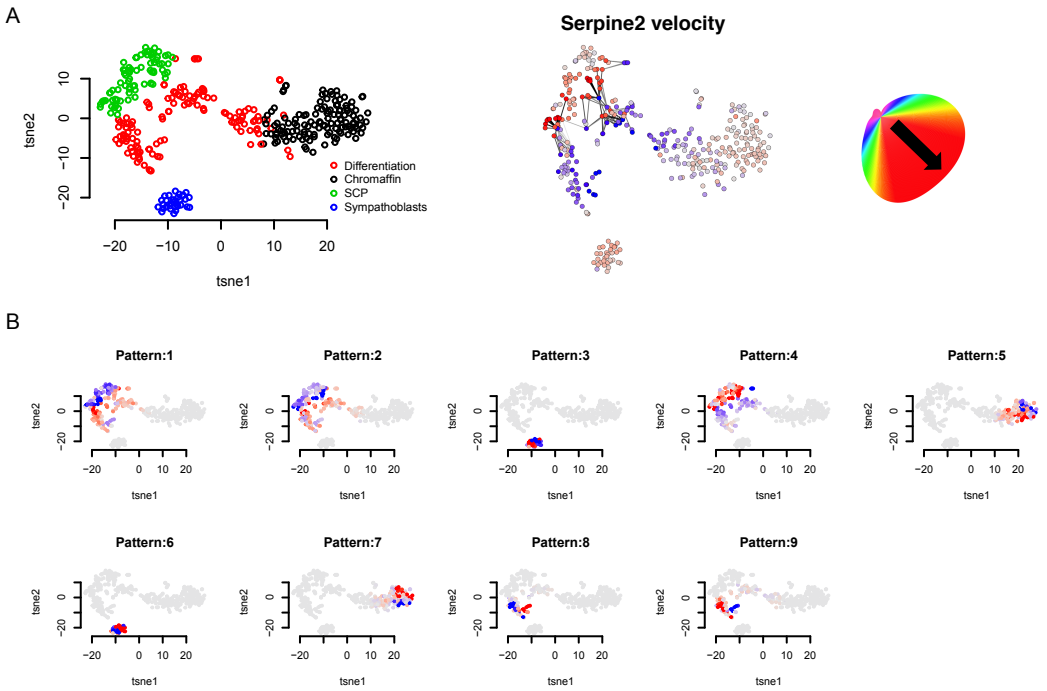


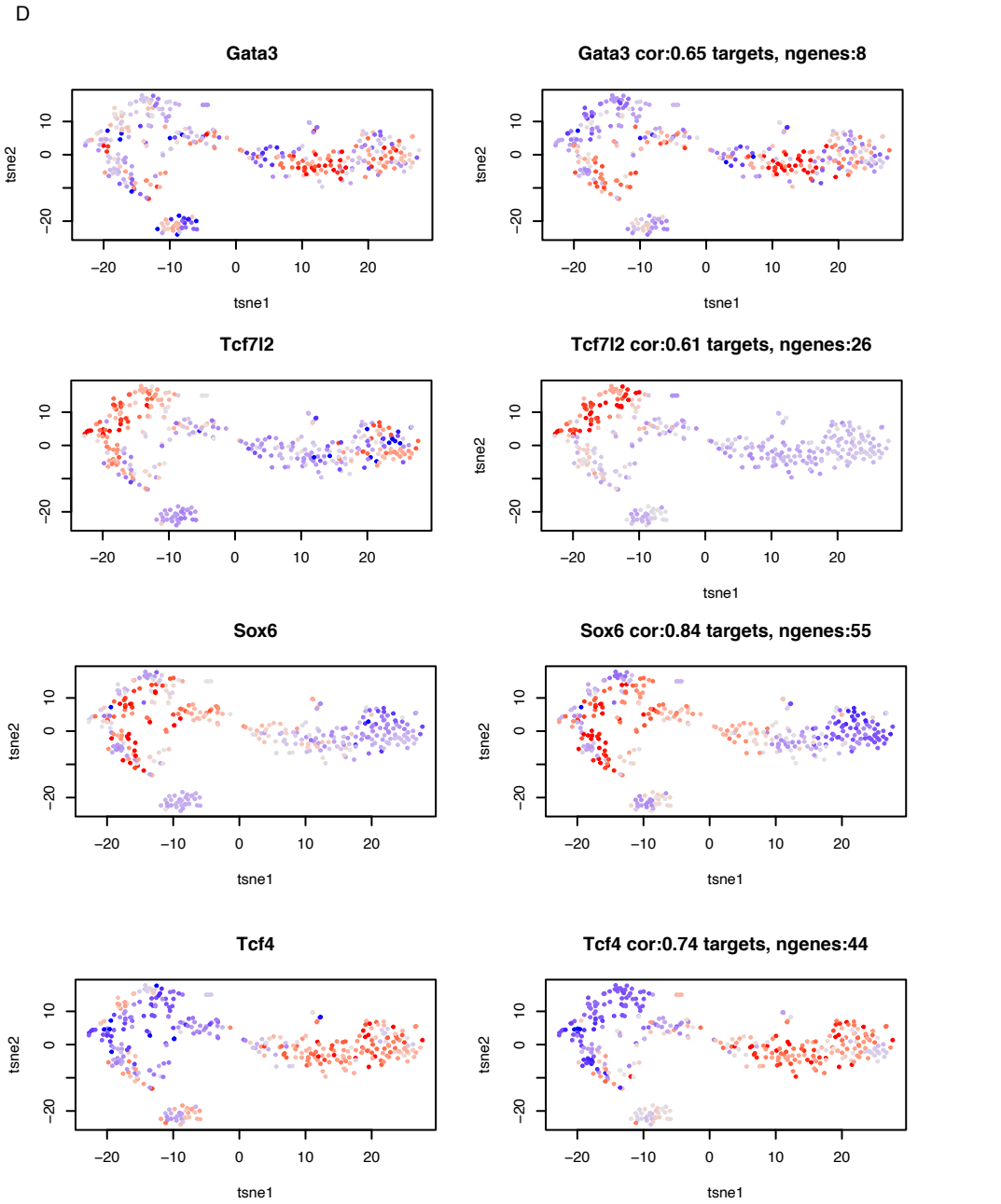


4. Identify gene regulatory network from velocity

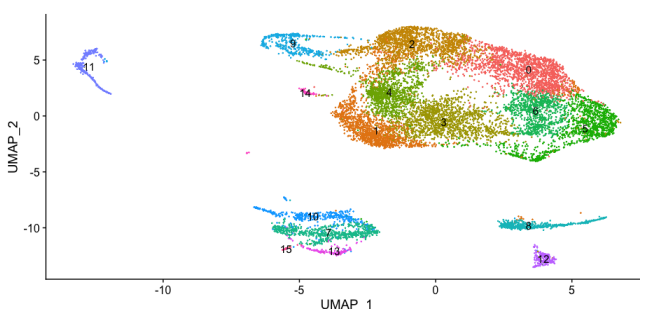


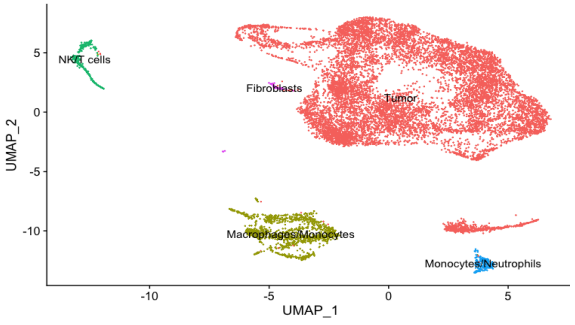

A
correlation
0.50
0.25
0.00
newVsVel
wholeVsVel
B
0.6
0.3
0.0
-0.3
wTF_wTarget
newTF_wTarget
velTF_wTarget
wTF_newTarget
newTF_newTarget
velTF_NewTarget
wTF_velTarget
newTF_velTarget
velTF_velTarget
C
CC_states
BRCA1_velocity
BRCA1_whole
BRCA1_new
BRCA1:Targets_velocity
BRCA1:Targets_whole
BRCA1:Targets_velocity
NF1_velocity
NF1_whole
NF1_new
NF1:Targets_velocity
NF1:Targets_whole
NF1:Targets_velocity
POLR2A_velocity
POLR2A_whole
POLR2A_new
POLR2A:Targets_velocity
POLR2A:Targets_whole
POLR2A:Targets_velocity
DEX_states
TEAD1_velocity
TEAD1_whole
TEAD1_new
TEAD1Target_velocity
TEAD1Target_whole
TEAD1Target_velocity
umap1
umap2


A
Serpine2 velocity
tsne2
tsne1
Differentiation
Chromaffin
SCP
Sympathoblasts
B
Pattern:1
Pattern:2
Pattern:3
Pattern:4
Pattern:5
Pattern:6
Pattern:7
Pattern:8
Pattern:9
C

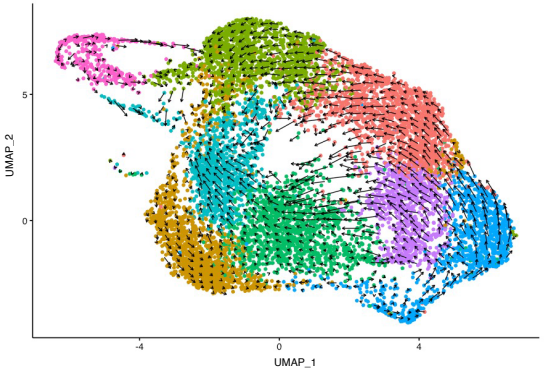

D
Gata3
Gata3 cor:0.65 targets, ngenes:8
Tcf7l2
Tcf7l2 cor:0.61 targets, ngenes:26
Sox6
Sox6 cor:0.84 targets, ngenes:55
Tcf4
Tcf4 cor:0.74 targets, ngenes:44
tsne1
tsne2

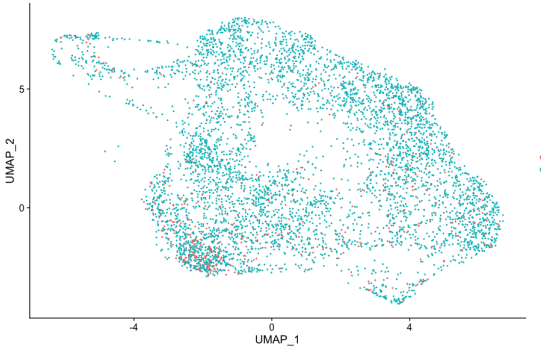
A


UMAP_2
UMAP_1
11
9
2
0
14
4
1
3
6
7
10
15
13
8
12

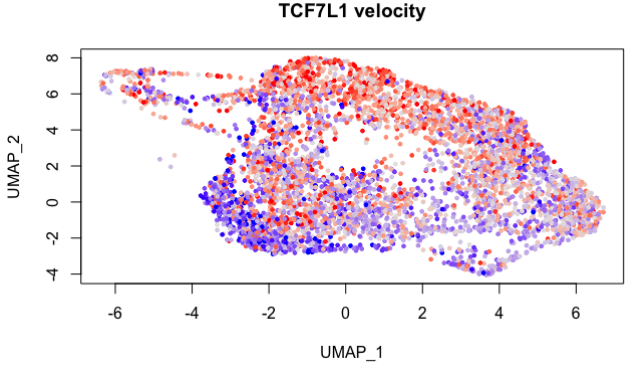

UMAP_2
UMAP_1
NK/T cells
Fibroblasts
Tumor
Macrophages/Monocytes
Monocytes/Neutrophils

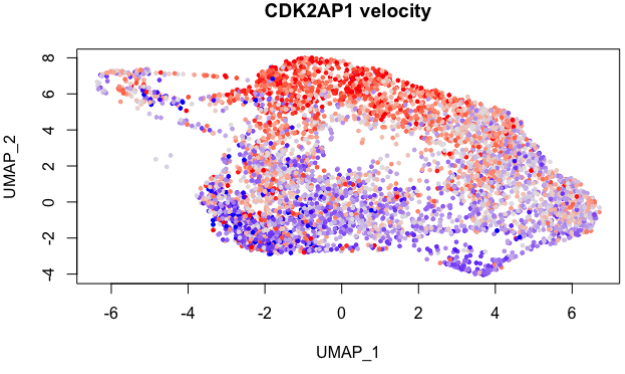
B


UMAP_2
UMAP_1


UMAP_2
UMAP_1

C


TCF7L1 velocity
UMAP_2
UMAP_1


CDK2AP1 velocity
UMAP_2
UMAP_1