VeloViz: RNA-velocity informed 2D embeddings for visualizing cellular trajectories

Lyla Atta[1,2,3], Jean Fan[1,2,4*]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

[2]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21211, USA

[3]Medical Scientist Training Program, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

[4]Department of Computer Science, Johns Hopkins University, Baltimore MD 21218, USA

[*]To whom correspondence should be addressed

Correspondence should be addressed to: Jean Fan (jeanfan@jhu.edu)

Key words: Bioinformatics, Computational biology, Gene expression, Single Cell, RNA velocity

Word count: 999

Figure count: 1

Table count: 0

0 Abstract

RNA velocity analysis can predict cell state changes from single cell transcriptomics data. To interpret these cell state changes as part of underlying cellular trajectories, current approaches rely on visualization with 2D embeddings derived from principal components, t-distributed stochastic neighbor embedding, among others. However, these 2D embeddings can yield different representations of the underlying trajectories, hindering the interpretation of cell state changes. To address this challenge, we developed VeloViz to create RNA-velocity-informed 2D embeddings. We show that by taking into consideration the predicted future transcriptional states from RNA velocity analysis, VeloViz can help ensure a more reliable representation of underlying cellular trajectories. VeloViz is available as an R package at https://github.com/JEFworks-Lab/veloviz.

1 Introduction

Single cell transcriptomics provide a static snapshot of transcriptional states for individual cells. The continuum of transcriptional states for cells along dynamic processes can be used to infer how cell states may change over time (Tritschler *et al.*, 2019; Saelens *et al.*, 2019). Notably, RNA velocity analysis can be applied to infer dynamics of gene expression and predict the future transcriptional state of a cell from single cell RNA-sequencing and imaging data (La Manno *et al.*, 2018; Xia *et al.*, 2019).

To interpret cell state changes from RNA velocity analysis, current approaches project the observed current and predicted future transcriptional states onto 2-dimensional (2D) embeddings to visualize the putative directed cellular trajectory (La Manno *et al.*, 2018; Zywitza *et al.*, 2018; Bastidas-Ponce *et al.*, 2019; Zhang *et al.*, 2019). Previously used 2D embeddings include those derived from principal

43 components (PC), t-distributed Stochastic Neighbor Embeddings (t-SNE), Uniform Manifold
44 Approximation and Projection (UMAP), or diffusion maps (Coifman *et al.*, 2005; Maaten and Hinton,
45 2008; McInnes *et al.*, 2018). However, these approaches can yield different representations of the
46 underlying trajectory. Furthermore, when intermediate cell states are not well represented, current 2D
47 embeddings may not capture global relationships between cell subpopulations, thereby further hindering
48 the interpretation of cell state changes (Kester and Oudenaarden, 2018; Weinreb *et al.*, 2018).
49
50 Here, we developed VeloViz to visualize cellular trajectories by incorporating information from RNA
51 velocity analysis. By taking into consideration cells' predicted future transcriptional states inferred from
52 RNA velocity analysis, VeloViz can help ensure that relationships between cell states are reflected in the
53 2D embedding, allowing for more reliable representation of underlying cellular trajectories.
54
55
56 **2 Method**
57
58 In order to create an RNA-velocity-informed 2D embedding, VeloViz uses each cell's current observed
59 and predicted future transcriptional states inferred from RNA velocity analysis to build a nearest neighbor
60 graph between cells in the population (Figure 1A). Briefly, VeloViz computes a cell-cell composite
61 distance between all cell pairs in the population (Fig 1A, Supplementary Information 1ii) and assigns
62 graph edges to the *k* neighboring cells with the smallest composite distances. Edges are then pruned based
63 on similarity thresholds (Supplementary Information 1iii). The resulting graph can be visualized as a 2D
64 embedding using force-directed layout algorithms (Fruchterman and Reingold, 1991).
65
66
67 **3 Results**

68

69 3.1 Comparing VeloViz to other embeddings

70 To evaluate the performance of VeloViz, we first assessed VeloViz's ability to capture trajectories of
71 simulated data representing cycling or branching trajectories (Supplementary Information 2i) and
72 compared to PC, t-SNE, UMAP, and diffusion map embeddings. We calculated a trajectory consistency
73 (TC) score (Supplementary Information 2ii., (Boggust et al., 2019)) where TC scores closer to 1 indicate
74 more accurate representations of the ground truth trajectory. Among evaluated trajectories, VeloViz
75 embeddings had consistently high TC scores (Supplementary Figure 1). Next, we used VeloViz to
76 visualize pancreatic endocrinogenesis single-cell RNA-sequencing (scRNA-seq) data, where cycling
77 ductal cells give rise to endocrine progenitor-precursor (EP) cells, which then differentiate into hormone
78 producing endocrine cell types (Alpha, Beta, Delta, and Epsilon cells) (Bastidas-Ponce *et al.*, 2019). We
79 observed that while all evaluated embeddings captured the trajectory of endocrine progenitors, VeloViz
80 was better able to capture the cycling structure of ductal cells (Supplementary Figure 2). VeloViz, UMAP,
81 and tSNE also captured the terminal branching differentiation into the different endocrine cell types, which
82 is not clear in PC or diffusion map. Overall, VeloViz is able to capture trajectories of diverse topologies.

83

84 3.2 Performance with incomplete trajectories

85 To evaluate the performance of VeloViz in visualizing trajectories with missing intermediate cell states,
86 we used simulated and real scRNA-seq data where some intermediate cells were removed, creating a
87 trajectory gap. Because t-SNE and UMAP preferentially preserve local cell-cell relationships, we expected
88 that these embeddings would result in two distinct clusters of cells before and after the simulated gap
89 (Kobak and Berens, 2019; Heiser and Lau, 2020). Therefore, in addition to TC scores, we calculated a
90 gap distance (Supplementary Information 2iii), which measures the distance in the 2D embedding space
91 between cells before and after the simulated gap in the trajectory. Embeddings that preserve the underlying
92 trajectory despite this simulated gap will have a smaller gap distance. Indeed, for the cycling trajectory

where cells corresponding to a segment of the cycle were removed, VeloViz was the only embedding able to preserve the cycling structure. Likewise, for branching trajectories with missing intermediates, only VeloViz and PCA were able to preserve the underlying topology while tSNE and UMAP split cells before and after the simulated gap into distinct clusters as expected (Supplementary Figure 3). TC scores were consistently higher and the gap distance smaller for VeloViz than with t-SNE, UMAP, and diffusion map. Likewise, for the pancreatic endocrinogenesis scRNA-seq data, we removed pre-endocrine cells and used cell latent time (Bergen et al., 2020) to identify cells before and after pre-endocrine cells in the developmental trajectory and to calculate gap distances (Supplementary Information 2iii). Notably, the transition from endocrine progenitors into terminal endocrine cell types was best captured by VeloViz. As expected, t-SNE and UMAP split ductal and endocrine progenitor cells from terminal endocrine cell types, which is reflected in the gap distances (Figure 1B-F). Overall, VeloViz is able to visualize a more reliable presentation of underlying trajectories even when intermediate cell states are missing.

Funding: This work was supported by the National Institutes of Health [T32GM136577 to L.A.]

Conflict of Interest: none declared.

References

Bastidas-Ponce,A. *et al.* (2019) Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, **146**.

Bergen,V. *et al.* (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*, 1–7.

Coifman,R.R. *et al.* (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.*, **102**, 7426–7431.

Fruchterman,T.M.J. and Reingold,E.M. (1991) Graph drawing by force-directed placement. *Softw. Pract. Exp.*, **21**, 1129–1164.

Kester,L. and Oudenaarden,A. van (2018) Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*, **23**, 166–179.

La Manno,G. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.

Maaten,L. van der and Hinton,G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.

McInnes,L. *et al.* (2018) UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, **3**, 861.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.

Tritschler,S. *et al.* (2019) Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, **146**.

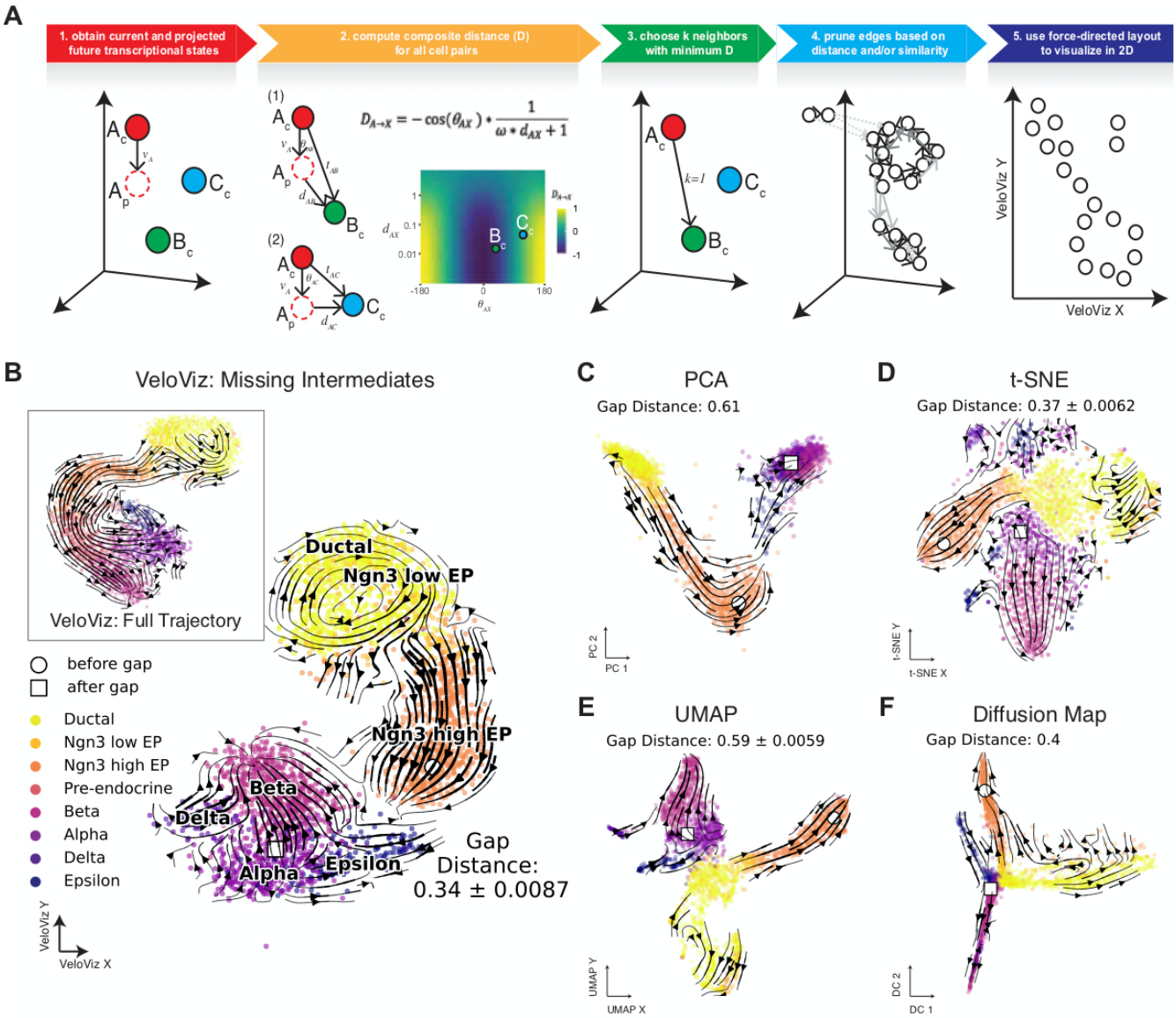
Weinreb,C. *et al.* (2018) Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci.*, **115**, E2467–E2476.

Xia,C. *et al.* (2019) Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci.*, **116**, 19490–19499.

Zhang,Q. *et al.* (2019) Landscape and Dynamics of Single Immune Cells in Hepatocellular Carcinoma. *Cell*, **179**, 829-845.e20.

Zywitza,V. *et al.* (2018) Single-Cell Transcriptomics Characterizes Cell Types in the Subventricular Zone and Uncovers Molecular Defects Impairing Adult Neurogenesis. *Cell Rep.*, **25**, 2457-2469.e8.

144 **Figures**



145

146 **Figure 1. VeloViz constructs RNA-velocity-informed 2D embeddings**. **A)** Workflow to create a
147 VeloViz 2D embedding: 1) Observed current (X_c) and predicted future (X_p) transcriptional cell states
148 inferred from RNA velocity are reduced into a common PC space; 2) composite distances (D) between all
149 cell pairs are computed. Composite distance from Cell A to Cell X ($D_{A\rightarrow X}$) takes into account the similarity
150 in transcriptional profiles (d_{AX}) between Cell X's observed current (X_c) and Cell A's predicted future
151 transcriptional state (A_p), and the cosine correlation between Cell A's RNA-velocity (v_A) and the change
152 vector (t_{AX}) representing a transition from Cell A's current state (A_c) to Cell X's current state (X_c). A
153 distance weight (ω) is used to adjust the relative importance of transcriptional similarity and cosine

154 correlation in the composite distance; 3) for each cell, graph edges are assigned to the *k* cells with the
155 minimum composite distances to create a graph. Edge weights are computed based on composite distances
156 as $weight_{AB} = max(D) - D_{AB}$; 4) edges assigned in 3. are removed (in grey, dashed) if they are above the
157 similarity and/or distance thresholds. Edge shade corresponds to edge weight computed based on
158 composite distance, with darker arrows representing edges with larger weights; 5) the resulting graph is
159 visualized as a 2D embedding using a force-directed graph layout. **B)** VeloViz 2D embedding visualizing
160 pancreatic endocrinogenesis with pre-endocrine intermediates removed creating a gap in the
161 developmental trajectory. Inset shows the VeloViz embedding of the full dataset. Cells are colored by cell
162 state annotations provided in (Bergen *et al.*, 2020). Arrows show the projection of velocities derived from
163 dynamical velocity modelling onto the VeloViz embeddings. Gap distances measure the median distance
164 in the 2D embedding between the 300 cells before and after pre-endocrine cells in the developmental
165 trajectory (Supplementary Information 2iii). White circle and square indicate the median coordinates of
166 cells before and after pre-endocrine cells in the developmental trajectory, respectively. **C-F)** 2D
167 embeddings visualizing pancreatic endocrinogenesis with removed pre-endocrine intermediates using
168 PCA, t-SNE, UMAP, and diffusion mapping, respectively with arrows showing the projection of velocities
169 derived from dynamical velocity modelling.