VeTra: a new trajectory inference tool based on RNA velocity

Guangzheng Weng[1], Junil Kim[2,3*] and Kyoung Jae Won[2,3*]

[1]Department of Biology, The bioinformatics Centre, University of Copenhagen, 2200 Copenhagen N, Denmark, [2]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen N, Denmark, [3]Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, Faculty of Health and Medical Sciences, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark

* To whom correspondence should be addressed to KJW (Tel: +45-3533-1419 ; Email: kyoung.won@bric.ku.dk) and JK (junil.kim@bric.ku.dk).

ABSTRACT

Motivation: Trajectory inference for single cell RNA sequencing (scRNAseq) data is a powerful approach to understand time-dependent cellular processes such as cell cycle and cellular development. However, it is still not easy to infer the trajectory precisely by which cells differentiate to multiple lineages or exhibit cyclic transitions. Recent development of RNA velocity provides a way to visualize cell state transition without a prior knowledge. Trajectory inference that utilizes the velocity information will be highly useful to understand cellular dynamics.

Results: We developed VeTra, a tool to infer the trajectories from scRNAseq data. Uniquely, VeTra can perform grouping of cells that are in the same stream of trajectory. For this, VeTra searches for weakly connected components of the directed graph obtained from RNA velocity. Therefore, VeTra makes it easy to define groups of cells from the origin and to the end of a certain trajectory. VeTra has been tested to infer the streams of cells for pancreatic development, neural development in hippocampus and cell cycle. VeTra is a useful tool to perform pseudo-time analysis from the start to the end of each group.

Keyword: single cell RNA sequencing, Trajectory Inference, RNA velocity, Weakly Connected Component, Multiple lineage differentiation, Cell cycle

INTRODUCTION

Trajectory analysis using single cell transcriptomics is useful to understand temporal transition of cell states. At present, more than 70 methods have been published for single-cell trajectory inference (TI) (Saelens et al. 2019). It is important for the trajectory analysis to precisely identify branching events and the direction of the trajectory. Also, grouping of cells based on the cell transition will be greatly useful for subsequent pseudo-time analysis.

TI methods integrated the information about the marker genes, the number of branching events, or a specified root cell to identify the direction, which are often provided by the users (Trapnell et al. 2014; Setty et al. 2016; Haghverdi et al. 2016). Automatic approaches to detect the branching structure were also developed using minimum spanning tree (Street et al. 2018) and graph-partitioning algorithm (Wolf et al. 2019).

Recently, RNA velocity was developed to visualize the direction of cell transition from RNA dynamics information (La Manno et al. 2018). RNA velocity enables visual inspection of groups of cells that belong to the same stream of cell state transition. Though useful, it is still not easy for a user to obtain the branched trajectory and the genes associated with them. It is more difficult when multiple branches share the same root.

We developed VeTra to group the cells that belong the same stream of trajectory. For this, VeTra searches for weakly connected components (WCC) in the directed graph of the cell transition information obtained from RNA velocity. As the results, VeTra uniquely has a function to select groups of cells that belong to the same transition stream compared with other TI algorithms. A cell can be assigned to multiple groups when located in the root of the branches. VeTra provides a flexible environment to obtain cell group, select trajectory of interest and enables pseudo-time analysis.

METHODS

VeTra reconstructs the pseudo-temporal orders of the cells based on the coordinate and the velocity vector of each cell in the low-dimensional space. The velocity vectors are estimated by extrapolating the spliced/unspliced ratio to the local neighboring cells (La Manno et al. 2018). Given velocity vectors (Figure 1A), VeTra reconstructs a directed graph. To pick the most appropriate cell to which the vector of the cell *i* is pointing, we collected the k (the number of neighbors) closest neighbor cells from the head of the vector of the cell *i* (Figure 1B). Among the k neighbor cells, we selected the cells whose vectors have similar direction using a cosine similarity criterion $\cos_{1ij} > 0.5$ between the cell *i* and the

neighbor cell *j* (Figure 1B). To obtain a directed graph, we removed the neighbors which are located behind the cell *i* along the stream.

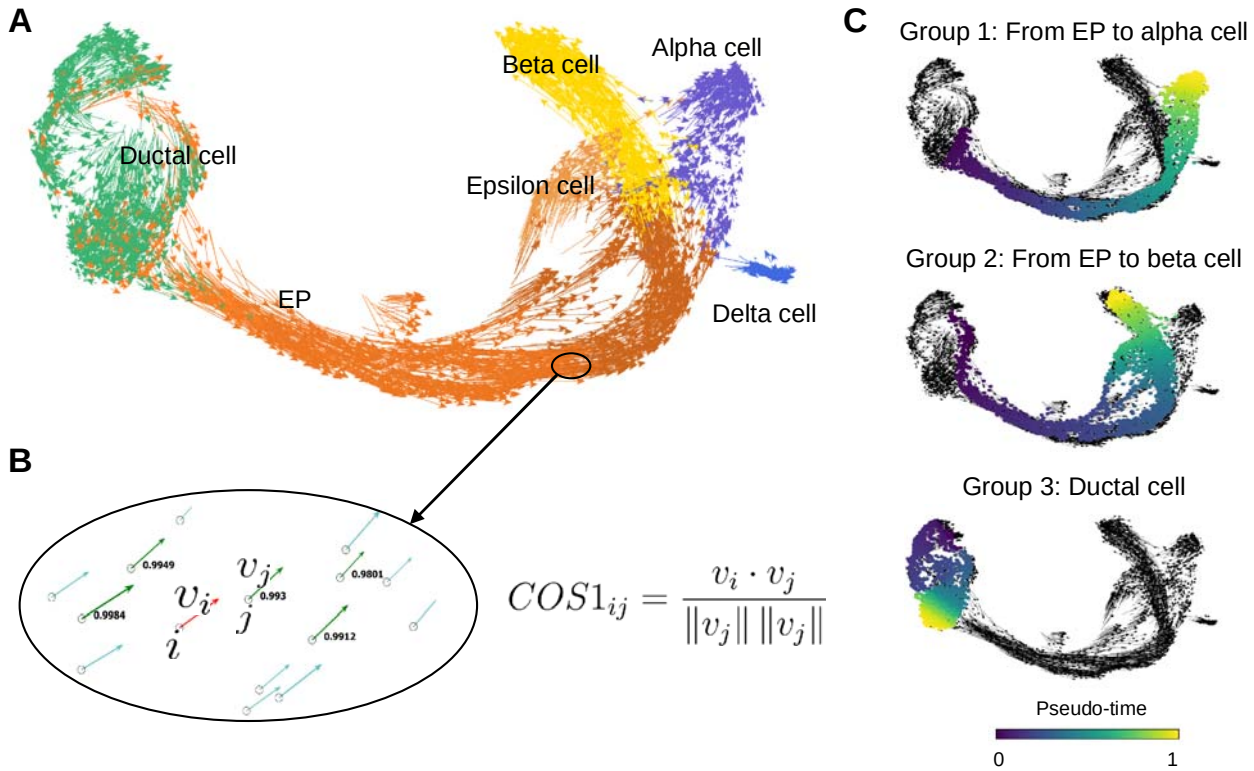

Fig. 1. VeTra reconstructs single-cell trajectories for multiple cell lineages. **A**. A 2D embedding plot of pancreatic development colored by cell types. **B**. Cosine similarity for searching closest neighbor cells pointing similar direction. **C**. Three trajectory groups inferred by VeTra.

We divided the directed graph into WCCs, which is defined by a maximum subgraph such that every node is reachable from every other node by an undirected path (Walker and Skiena 1992). Hierarchical clustering for the WCC will provide foundation for grouping. The distance between two WCCs is defined by the maximum distance of all the closest pairs of cells from the smaller WCC. For the distance between cells, we calculated the Euclidean distance in the four-dimensional space (two dimensions from the reduced dimensions of gene expression and the other two dimensions from the velocity vector). To obtain full trajectory (from the root to the branch), we extended the members of the clusters if the members located nearby from each cell in the group with $\cos_{1ij} > 0.7$. We obtained the pseudo-time ordering by projecting the member cells onto the principal curve (Hastie and Stuetzle 1989).

RESULTS

To confirm that VeTra can cluster multifurcated trajectories, we applied VeTra to two scRNAseq datasets for the pancreatic development (Bastidas-Ponce et al. 2019; Bergen et al. 2020) and neural lineages in the hippocampus (La Manno et al. 2018). From the transcriptome of pancreatic development, VeTra identified three trajectory clusters of i) alpha cell differentiation from endocrine progenitors (EP), ii) beta/epsilon differentiation from EP, and iii) ductal cells (Figure 1D). The first two clusters were commonly originated from EPs and bifurcate into different lineages. We also obtained small clusters for delta cell lineage (Figure 2). With the dataset for neural lineages, VeTra identified all five lineages CA1-subiculum, CA2-3-4, granule, astrocytes, oligodendrocyte progenitors (Figure 3). CA2-3-4 cells and granules share the same neuroblastic origin and astrocytes and oligodendrocyte progenitors share the same intermediate progenitor origin. These results showed that VeTra automatically captures the branching structure.

We also applied VeTra to the scRNAseq dataset for cell cycle progression (Xia et al. 2019). VeTra identified a single cycle-shaped trajectory cluster, starting at G1 phase, all the way through G1/S, G2/M, and M phase, and finally going back to G1 phase (Figure 4).

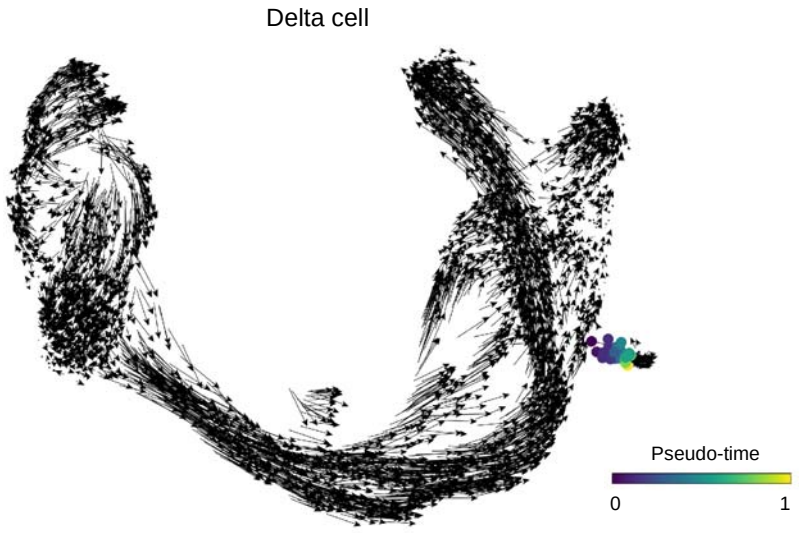

Figure 2. VeTra identifies small clusters for delta cell lineage from a scRNAseq dataset of pancreatic cells.

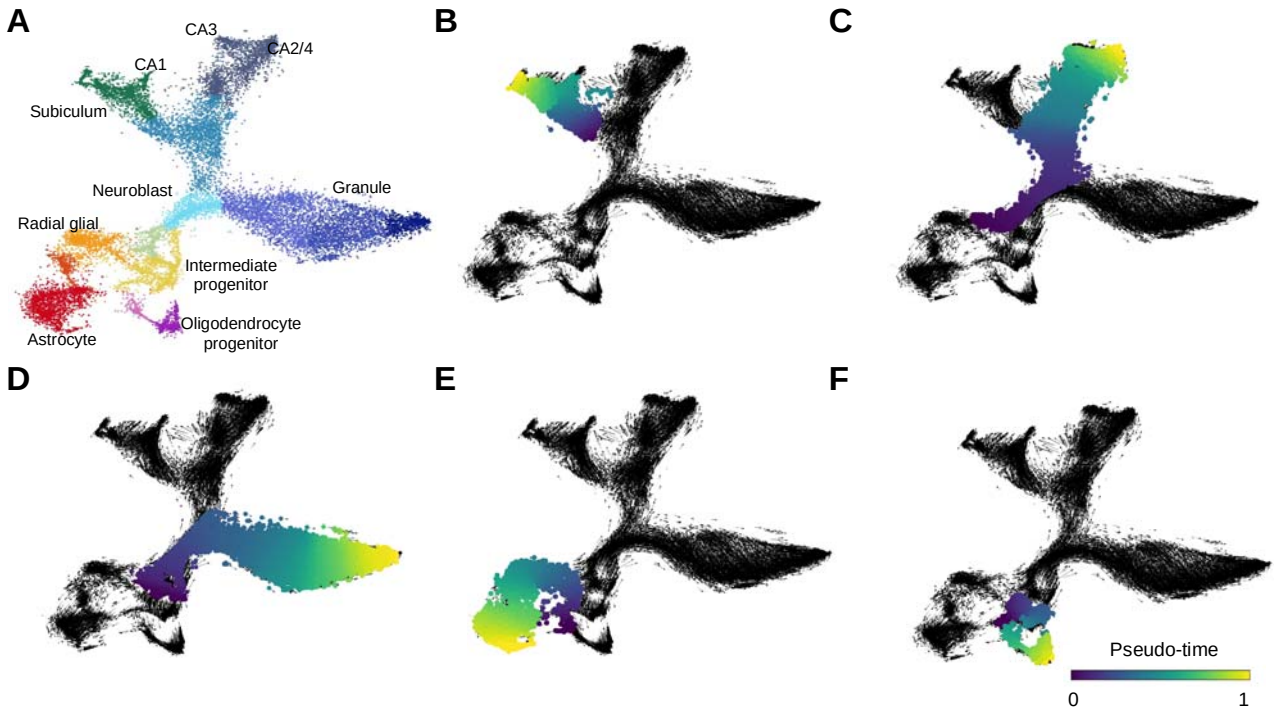

Figure 3. VeTra successfully identifies five neural cell lineages from a scRNAseq dataset of hippocampus. **A**. A 2D embedding plot of neural cell lineages colored by cell types. 2D embedding plots of five trajectory groups of neural cell lineages CA1-subiculum (**B**), CA2-3-4 (**C**), granule (**D**), astrocyte (**E**), oligodendrocyte progenitor (**F**), inferred by VeTra.

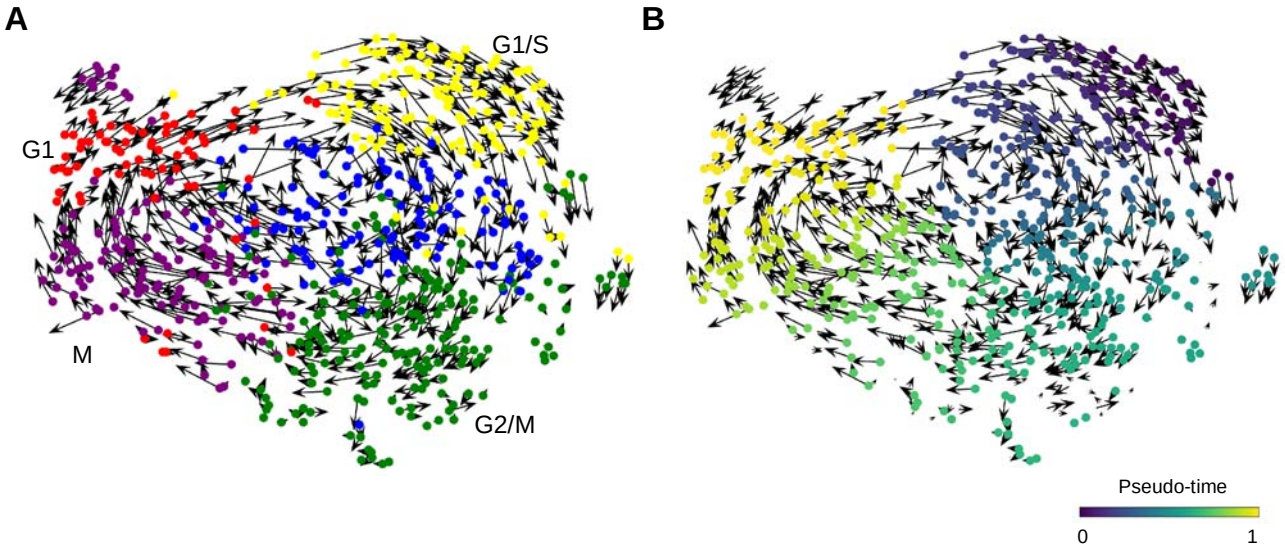

Figure 4. VeTra capture a single trajectory group with cyclic topology from a scRNAseq dataset for cell cycle progression. 2D embedding plots of cells colored by cell cycle clusters (**A**) and pseudo-time inferred by VeTra (**B**)

DISCUSSION

AVAILABILITY

Vetra is available at https://github.com/wgzgithub/VeTra

FUNDING

The Novo Nordisk Foundation Center for Stem Cell Biology is supported by a Novo Nordisk Foundation grant No. NNF17CC0027852. KJW is supported by the Lundbeck Foundation's Ascending Investigator grant R313-2019-421 and the NIH grant (R01 DK106027).

Competing interests

The authors declare no competing interests.

REFERENCES

Bastidas-Ponce A, Tritschler S, Dony L, Scheibner K, Tarquis-Medina M, Salinno C, Schirge S, Burtscher I, Böttcher A, Theis FJ, et al. 2019. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Dev* **146**: dev173849.

Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*.

Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**: 845–8.

Hastie T, Stuetzle W. 1989. Principal curves. *J Am Stat Assoc* **84**: 502–516.

La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriti ME, Lönnerberg P, Furlan A, et al. 2018. RNA velocity of single cells. *Nature* **560**: 494–498.

Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**: 547–554.

Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'Er D. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**: 637–645.

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. 2018. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**: 477.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**: 381–6.

Walker R, Skiena S. 1992. Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. *Math Gaz*.

Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**: 59.

Xia C, Fan J, Emanuel G, Hao J, Zhuang X. 2019. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci U S A* **116**: 19490–19499.