bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

VIA: Generalized and scalable trajectory inference in single-cell omics data

Shobana V. Stassen[1], Gwinky G. K. Yip[1], Kenneth K. Y. Wong[1,3], Joshua W. K. Ho[2,4] and Kevin K. Tsia[1,3]

[1]Department of Electrical & Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong
[2]School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong
[3]Advanced Biomedical Instrumentation Centre, Hong Kong Science Park, Shatin, New Territories, Hong Kong
[4]Laboratory of Data Discovery for Health, Hong Kong Science Park, Shatin, New Territories, Hong Kong

Abstract

Inferring cellular trajectories using a variety of omic data is a critical task in single-cell data science. However, prediction and thus biologically meaningful discovery of cell fates are challenged by the sheer size of single-cell data, diverse omic data types, and their complex data topologies. We present VIA, a scalable trajectory inference algorithm that uses lazy-teleporting random walks to accurately reconstruct complex cellular trajectories beyond tree-like pathways (e.g. cyclic or disconnected structures), and to discover less populous lineages or those otherwise obscured in other methods. VIA outperforms existing algorithms in recapitulating cell fates/lineages, and also mitigates loss of global connectivity information in large datasets beyond a million cells. Furthermore, VIA demonstrates versatility by distilling cellular trajectories in single-cell transcriptomic, epigenomic, proteomic and morphological data – showing new promise in scalable, multifaceted single-cell analysis to explore novel biological processes.

Introduction

Single-cell omics data captures snapshots of cells that catalog cell types and molecular states with high precision. These high-content single-cell readouts can be harnessed to model evolving cellular heterogeneity and track dynamical changes of cell fates in tissue, tumour, and cell population. However, current computational methods face four critical challenges. First, it remains difficult to accurately reconstruct high-resolution cell trajectories and detect cell fates embedded within them. Even the few algorithms which automate cell fate detection (e.g., SlingShot[1] and Palantir[2]) exhibit low sensitivity and are highly susceptible to changes in input parameters. Second, current trajectory inference (TI) methods predominantly work well on tree-like trajectories (e.g. Slingshot, Monocle2[3]), but lack the generalisability to infer disconnected, cyclic or hybrid topologies without imposing restrictions on transitions and causality[4]. Third, the growing scale of single-cell data, notably cell atlases of whole organisms[6,7], embryos[8,9] and human organs[10], exceeds the existing TI capacity, not just in runtime and memory, but in preserving global connectivity, which is often lost after extensive dimension reduction or subsampling. Fourth, fueling the advance in single-cell technologies is the ongoing pursuit to understand cellular heterogeneity from a broader perspective beyond transcriptomics. However, the applicability of TI to a broader spectrum of single-cell data has yet to be fully exploited.

To overcome these recurring challenges, we present VIA, a graph-based TI algorithm that uses a new strategy to compute pseudotime, and reconstruct cell lineages based on lazy-teleporting random walks integrated with Markov chain Monte Carlo (MCMC) refinement. VIA relaxes common constraints on traversing the graph by allowing cyclic and temporally reversed movements, and thus robustly detects cell fates involving complex transitions that are otherwise obscured in other methods. VIA outperforms popular TI algorithms in terms of capturing cellular trajectories not limited to multi-furcations and trees, but also disconnected and cyclic topologies (**Supplementary Fig. S1)**. Compared to existing TI methods,

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

VIA is highly scalable with respect to number of cells (10^2 to $>10^6$ cells) and features, without requiring extensive dimensionality reduction or subsampling which compromise global information. We demonstrate VIA's accuracy, scalability, topological-generalizability and multi-omic versatility across multiple modalities by investigating 10 simulated and experimental datasets (**Supplementary Table S1**), ranging from single-cell RNA-sequencing (scRNA-seq), single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq), multi-omics integration, to mass and imaging cytometry.

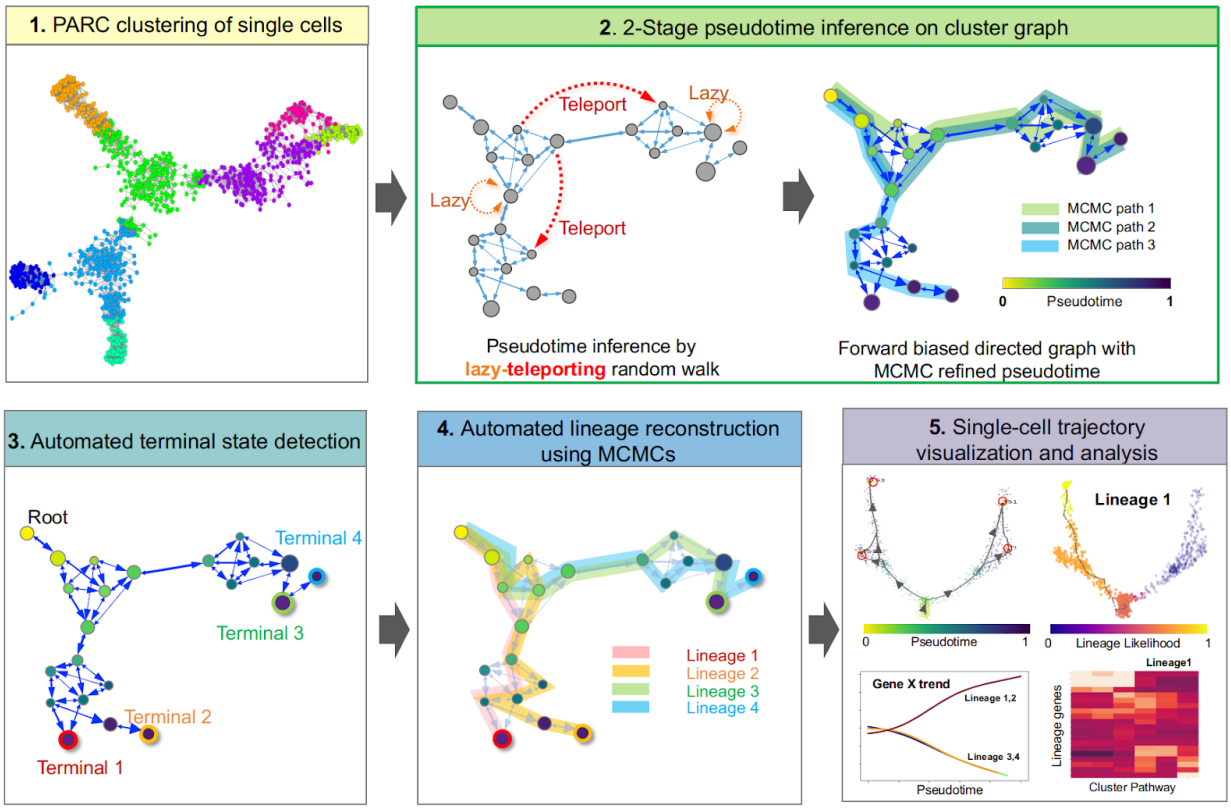


Figure 1. General workflow of VIA algorithm. Step 1: Single-cell level graph is clustered such that each node represents a cluster of single cells (computed by our clustering algorithm PARC[11]). The resulting cluster graph forms the basis for subsequent random walks. **Step 2:** 2-stage pseudotime computation: (i) The pseudotime (relative to a user defined start cell) is first computed by the expected hitting time for a lazy-teleporting random walk along an undirected graph. At each step, the walk (with small probability) can remain (orange arrows) or teleport (red arrows) to any other state. (ii) Edges are then forward biased based on the expected hitting time (See forward biased edges illustrated as the imbalance of double-arrowhead size). The pseudotime is further refined on the directed graph by running Markov chain Monte Carlo (MCMC) simulations (See 3 highlighted paths starting at root). **Step 3:** Consensus vote on terminal states based on vertex connectivity properties of the directed graph. **Step 4**: lineage likelihoods computed as the visitation frequency under lazy-teleporting MCMC simulations. **Step 5**: visualization that combines network topology and single-cell level pseudotime/lineage probability properties onto an embedding using GAMs, as well as unsupervised downstream analysis (e.g. gene expression trend along pseudotime for each lineage).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

Algorithm

VIA first represents the single-cell data as a cluster graph (i.e. each node is a cluster of single cells), computed by our recently developed data-driven community-detection algorithm, PARC, which allows scalable clustering whilst preserving global properties of the topology needed for accurate TI[11] (**Step 1 in Fig. 1)**. The cell fates and their lineage pathways are then computed by a two-stage probabilistic method, which is the key algorithmic contribution of this work (**Step 2 in Fig. 1**, see **Methods**). In the first stage, VIA models the cellular process as a modified random walk that allows degrees of *laziness* (remaining at a node/state) and *teleportation* (jumping to any other node/state) with pre-defined probabilities. The pseudotime, and thus the graph directionality, can be computed based on the theoretical hitting times of nodes (See the theory and derivation in **Methods and Supplementary Note 2**). The lazy-teleporting behavior prevents the expected hitting time from converging to a local distribution in the graph as otherwise occurs in regular random walks, especially when the sample size grows[12]. More specifically, the laziness and teleportation factors regulate the weights given to each eigenvector-value pair in the expected hitting time formulation such that the stationary distribution (given by the local-node degree-properties in regular walks) does not overwhelm the global information provided by other 'eigen-pairs'. Moreover, the computation does not require subsetting the first k eigenvectors (bypassing the need for the user to select a suitable threshold or subset of eigenvectors) since the dimensionality is not on the order of number of cells, but equal to the number of clusters. Hence all eigenvalue-eigenvector pairs can be incorporated without causing a bottleneck in runtime. Consequently in VIA, the modified walk on a cluster-graph not only enables scalable pseudotime computation for large datasets in terms of runtime, but also preserves information about the global neighborhood relationships within the graph. In the second stage of Step 2, VIA infers the directionality of the graph by biasing the edge-weights with the initial pseudotime computations, and refines the pseudotime through MCMC simulations. Next **(Step 3 in Fig. 1),** the MCMC-refined graph-edges of the lazy-teleporting random walk enable accurate predictions of terminal cell fates through a consensus vote of various vertex connectivity properties derived from the directed graph. The cell fate predictions obtained using this approach are more robust to changes in input data and parameters compared to other TI methods (**Supplementary Fig. S1 and Fig. S16)**. Trajectories towards identified terminal states are resolved using lazy-teleporting MCMC simulations (**Step 4 in Fig. 1**). The probabilistic approach and relaxation of edge constraints allowed by VIA in computing differentiation pathways and pseudotime enables greater sensitivity to cell fates and complex trajectories, and makes allowances for asynchrony in differentiation processes by avoiding prematurely imposing constraints on node-to-node mobility. Other methods resort to constraints such as reducing the graph to a tree, imposing unidirectionality by thresholding edges based on pseudotime directionality, removing outgoing edges from terminal states[13,2] and computing shortest paths for pseudotime[2,1]. VIA's probabilistic approach to graph-traversal allows it to infer cell fates when the underlying data spans combinations of multifurcating trees and cyclic/disconnected topologies - topologies and hence lineages often obscured in existing TI methods (**Supplementary Fig. S1**). Together, these four steps facilitate holistic topological visualization of TI on the single-cell level (e.g. using UMAP or PHATE embeddings[14,15]) and other data-driven downstream analyses such as recovering gene expression trends (**Methods**). (**Step 5 in Fig. 1**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

VIA accurately infers trajectories in diverse scRNA-seq data

VIA recapitulates differentiation topologies and identifies elusive cell fates across a wide range of transcriptomic data. We first showcase the ability of VIA to explore large single-cell transcriptomic datasets by employing the 1.3-million-cell mouse organogenesis cell atlas (MOCA)[8]. While this dataset is inaccessible to most TI methods from a runtime and memory perspective, VIA can efficiently resolve the underlying developmental heterogeneity, including 9 major trajectories (**Fig. 2a,b**) with a runtime of ~40 min, compared to the next fastest method which has a runtime of at least 4 hours[2] (**Supplementary Table S3**). VIA preserves wider neighborhood information and reveals a globally connected topology of MOCA which is otherwise lost in the previous method. Broadly speaking, the overall cluster graph of VIA consists of three main branches that concur with the known developmental process at early organogenesis.[16] (**Fig. 2a)**. It starts from the root stem which has a high concentration of E9.5 early epithelial cells made of multiple sub-trajectories (e.g. epidermis, nose and foregut/hindgut epithelial cells derived from the ectoderm and endoderm). The stem is connected to two distinct lineages: 1) mesenchymal cells originated from the mesoderm which arises from interactions between the ectoderm and endoderm[17] and 2) neural tube/crest cells derived from neurulation when the ectoderm folds inwards[1]. The sparsity of early cells (only ~8% are E9.5) and the absence of earlier ancestral cells make it particularly challenging to capture the simultaneous development of trajectories. However, the overall pseudotime structure presented by VIA is reasonable. For instance, at the junction of the epithelial-mesenchymal branch, we find early mesenchymal cells from E9.5-E10.5. Cells from later mesenchymal developmental stages (e.g. myocytes from E12.5- E13.5) reside at the leaves of branches. Similarly, at the junction of epithelial-neural tube, we find dorsal tube neural cells and notochord plate cells which are predominantly from E9.5-E10.5 and more developed neural cells at the tips (e.g. excitatory and inhibitory neurons from E12.5-E13.5). VIA also places the other dispersed groups of trajectories (e.g. endothelial, hematopoietic) in biologically relevant neighborhoods (**Supplementary Notes 3, Supplementary Fig. S11**). While VIA's connected topology offers a coarse-grained holistic view, it does not compromise the ability to delineate individual lineage pathways (consistent with those found by Cao et al.,[8]) as shown in **Fig. 2c and Supplementary Fig. S11.** TI using VIA uniquely preserves both the global and local structures of the data and is thus particularly favorable for biological exploration involving large datasets, especially for comparative studies involving cell atlases[19]. Whilst manifold-learning methods are often used to extensively reduce dimensionality to mitigate the computational burden of large single-cell datasets, they tend to incur loss of global information and be sensitive to input parameters. VIA is sufficiently scalable to bypass such a step, and therefore retains a higher degree of neighborhood information when mapping large datasets. This is in contrast to Monocle3's[8] UMAP-reduced inputs that reveal different disconnected super-groups and fluctuating connectivity depending on input parameters (see **Supplementary Fig. S12-15** for the biologically consistent structures proposed by VIA across a range of parameters compared to the contradicting cell super groups and connectivity suggested by a UMAP based TI interpretation**).**

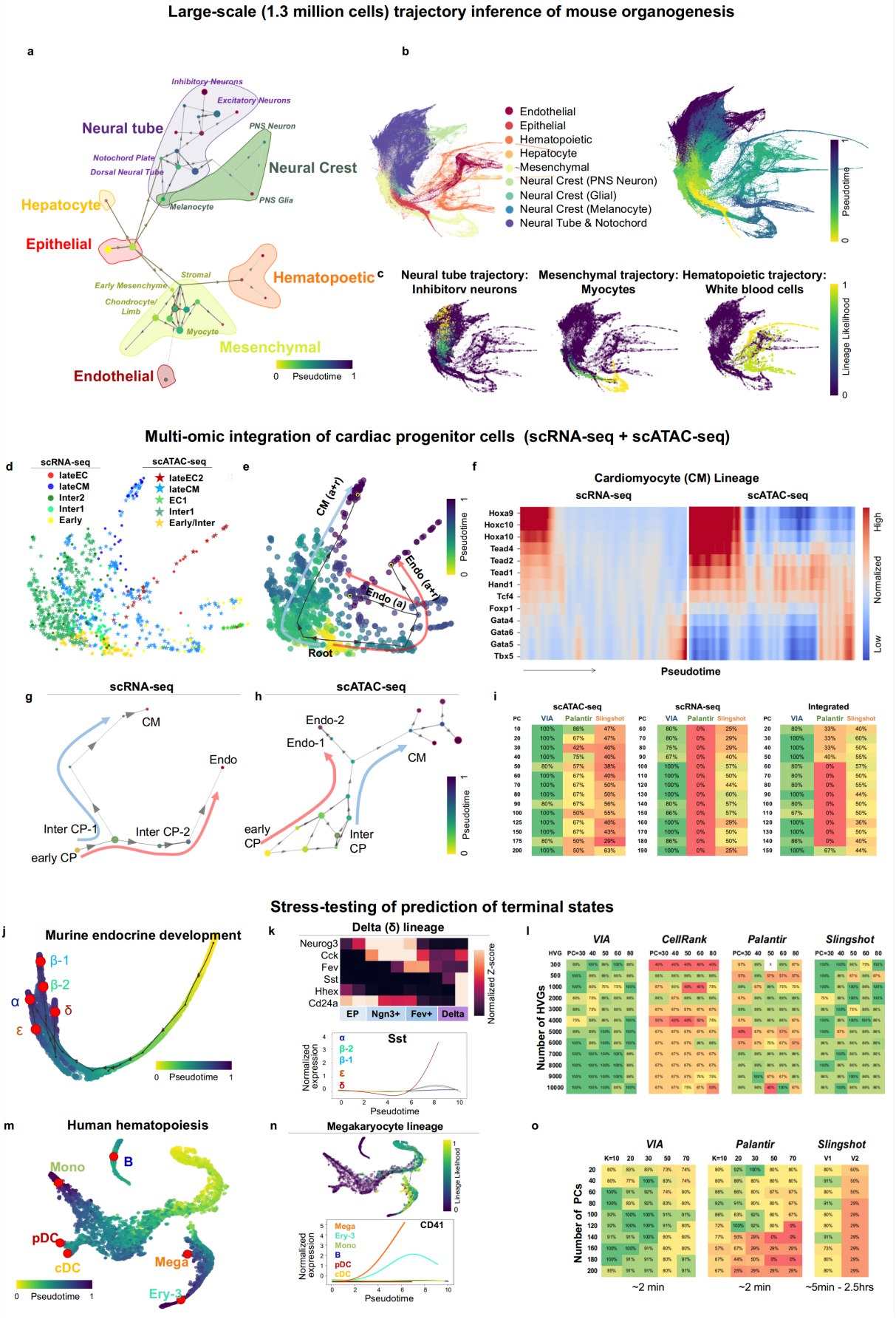
We next demonstrated the applicability of VIA in single-cell multi-omics analysis by inferring murine *Isl1*+ cardiac progenitor cell (CPC) transition states using both single-cell transcriptomic and chromatin accessibility information[20] (**Fig. 2d-i**). VIA consistently uncovers the bifurcating lineages towards the

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

endothelial and cardiomyocyte fates based on the scRNA-seq, scATAC-seq datasets and their data integration **(**see **Methods** for data integration). Other methods such as Palantir and Slingshot, that are also applicable to non-transcriptomic data, fail to uncover the two main lineages in the individual as well as the more challenging integrated multi-omic data. They typically only detect one of the two lineages and instead falsely detect several intermediate and early stages as final cell fates (**see Fig. 2i** for prediction accuracy). PAGA does not offer automated cell fate prediction and is therefore not benchmarked for this dataset. VIA detects lineage pathways in both the scRNA-seq and scATAC-seq that can be used to interpret relationships between transcription factor dynamics and gene expression in an unsupervised manner. VIA automatically generates a pseudotemporal ordering of cells (without requiring manual selection of relevant cells as done in Jia et al.[20]) along respective lineages and their marker-TF pairs (**see Fig. 2f and Supplementary Fig. S8e)**. The highlighted gene and TF pairs in the cardiac lineage show a strong correlation between expression and accessibility of *Gata* and Homeobox *Hox* genes which are known to be related to the regulation of cardiomyocyte proliferation[23,24,25]. VIA's reliable performance against user-reconfiguration (choice of components, individual or integrated omic data) suggests it can be used for transferable interpretation between scRNA-seq and chromatin accessibility data.

We further tested VIA on a wider scope of (small-to mid-sized) scRNA-seq datasets, including B-cell differentiation[26], hematopoiesis[2,27], embryonic stem (ES) cell differentiation in embryoid bodies[15], and endocrine differentiation (~10^2 - 10^4 cells). By comparing VIA with top-performing and popular TI algorithms, e.g. PAGA[28], Palintir, SlingShot and CellRank[13] (See **Methods**, and **Supplementary Fig. S1-7** for full analysis), we showed that VIA consistently outperforms other methods in terms of both runtime (in some cases by several magnitudes **see Supplementary Table S3 for runtime comparison**), and more robust and accurate lineage prediction across a wide range of pre-processing and algorithmic parameters. VIA's relaxation of graph traversal to permit cyclic sub-paths **(see Supplementary Fig. S1)** and movements that are temporally reversed, augments its sensitivity to lineages. Notably, VIA more consistently across a wide range of input parameter choice identified less populous lineages that were at best detected by other methods for a narrow sweet spot of parameters. For example, VIA reliably delineates the megakaryocyte, conventional and plasmacytoid dendritic cell (cDC and pDC) lineages in human hematopoiesis (**Fig. 2m-o, Supplementary Fig. S3-4** for pseudotime and graph-topological gene trends for all lineages); and Delta cells (3%) during the endocrine progenitor cells differentiation (**Fig. 2j-l, Supplementary Fig. S6** for pseudotime and topological gene trends for all lineages), as evidenced by the corresponding gene-expression trend analysis and parameter stress tests. Interestingly, we find that VIA often detects 2 Beta cell subpopulations **(Supplementary Fig. S6b,d,f)** that express typical Beta markers like *Dlk1, Pdx1***,** but differ in their expression of *Ins1* and *Ins2* **(Supplementary Fig. S6d)**. Such a Beta cell heterogeneity[29,30], whereby the immature Beta-2 population strongly expresses *Ins2*, and weakly expresses *Ins1*, and the mature Beta-1 population expresses both types of *Ins*[30], can also be reconciled based on the position of the Beta-2 cluster on the VIA graph **(Supplementary Fig. S6f).**

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 2 VIA accurately infers trajectories in diverse scRNA-seq datasets. (a) VIA cluster-graph trajectory where nodes are colored by pseudotime, and branches are shaded according to major lineages of 1.3-million-cell mouse organogenesis cell atlas (MOCA). The VIA analysis (which is independent of the choice of visualization) produces a connected structure with linkages between some of the major cell types that have a tendency to become segregated in a UMAP based TI analysis (see **Supplementary Fig. S12-15**). The stem (root) branch consists of epithelial cells derived from ectoderm and endoderm, leading to two main branches: 1) the mesenchymal and 2) the neural tube and neural crest. Other major groups are placed in the biologically relevant neighborhoods, such as the adjacencies between hepatocyte and epithelial trajectories; the neural crest (comprising glial cells and PNS neurons) and the neural tube; as well as the links between early mesenchyme with both the hematopoietic cells and the endothelial cells (**see Supplementary Note 3**). **(b)** (Left) Single-cell PHATE embedding colored by major cell groups. (Right) Single-cell PHATE embedding colored by VIA pseudotime. **(c)** Lineage pathways and probabilities of neuronal, myocyte and WBC lineages (**see Supplementary Fig. S6 for other lineages**). **(d)** scRNA-seq and scATAC-seq data of *Isl1*+ cardiac progenitors (CPs) integrated using Seurat3 before VIA TI analysis and PHATE visualization. Cells are colored by annotated cell-type and experimental modality **(e)** Cells are colored by VIA pseudotime with the VIA-inferred trajectory towards endothelial and myocyte lineages projected on top. **(f)** Marker gene expression and chromatin accessibility for gene-TF pairs along pseudotime axis for Cardiomyocyte lineage **(g)** VIA-graph trajectory with nodes colored by pseudotime shows bifurcation to endothelial and myocyte cells in scRNA-seq cells **(h)** scATAC-seq of *Isl1*+ CPs: VIA-graph again shows bifurcation after intermediate CP stage. **(i)** Lineage prediction accuracy (F1-score) for methods that offer automated lineage detection and are not limited to transcriptomic data. **(k)** Pancreatic Islets: Colored by VIA pseudotime with detected terminal states shown in red and annotated based on known cell type as Alpha, Beta-1, Beta-2, Delta and Epsilon lineages where Beta-2 is *Ins1*^{low}*Ins2*+ Beta subtype (**Supplementary Fig. S7**). **(l)** VIA inferred cluster-level pathway shows gene regulation along endocrine progenitor (EP) to Delta lineage specification (top) and Sst gene-expression trend shows rise of *Sst* in Delta lineage (**See Supplementary Fig. S7 for remaining**). **(m)** Prediction Accuracy of the 4 major endocrine cell types when varying the number of HVGs selected in pre-processing, and the number of PCs. **(n)** Human *CD34*+ hematopoiesis with 6 detected cell fates annotated **(o)** lineage pathway and gene-pseudotime trend shown for the *CD41* Megakaryocytic cells (**see Supplementary Fig. S3 for other lineages**). **(p)** Prediction accuracy of 6 cell fates when varying number of K (nearest neighbors) and PCs. Note Slingshot on default mode ("V2") uses GMM clustering and "V1" uses K-means clustering (allowing for over-clustering K=15, to increase sensitivity). Runtime of each method is also highlighted below the chart.

VIA enables multi-omic analysis beyond transcriptomic data

Broad applicability of TI beyond transcriptomic analysis is increasingly critical, but existing methods have limitations contending with the disparity in the data structure (e.g. sparsity and dimensionality) across a variety of single-cell data types and oftentimes are designed with a view to only handling transcriptomic data (e.g. methods using RNA velocity to infer directionality).

First, we employ VIA to analyze human scATAC-seq profiles (from *CD34*+ human bone marrow) (**Fig. 3a**), and find that the continuous landscape of hematopoiesis generally mirrors the scRNA-seq human hematopoietic data (**Fig. 2c**). The intrinsic sparsity of scATAC-seq data poses a challenge that can be alleviated by choice of pre-processing pipelines, and we see that VIA consistently predicts the expected hierarchy of furcations that leads to the lymphoid, myeloid and erythroid lineages for two commonly accepted pre-processing protocols[31,27] (**Methods**). This again holds across a wide range of input parameters, as shown by the changes in the accessibility of TF motifs associated with known regulators, e.g. *Gata1* (erythroid), *Cebpd* (myeloid) **(Fig 3b-d, Supplementary Fig. S7).**

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We next investigated whether VIA can cope with a significant drop in data dimensionality (10-100), as often presented in flow/mass cytometry data, and still delineate continuous biological processes. We run VIA on a time-series mass cytometry data (28 antibodies, 90K cells) capturing murine embryonic stem cells (ESCs) differentiation toward mesoderm cells (Day 0 - Day 11)[32]. Unlike previous analysis[32] of the same data which required chronological labels to visualize the developmental hierarchy, we ran VIA without such supervised adjustments and accurately captured the sequential development. VIA computed the trajectories with faster runtime (running in 2 minutes versus Slingshot which required 6 hours **see Table S3**), detecting 3 terminal states corresponding to cells in the final developmental stages: 2 corresponding to the main region of Day 10-11 (marked by *Pdgfra*, *Cd44* and *Gata4* expressions), and a small population of Day 10-11 cells expressing EpCAM, which are otherwise obscured in other methods (e.g. Palantir, Slingshot), especially the small EpCAM population (~0.5% of cells) **(Fig 3e-h, Fig. S9e,f)**.

Finally we tested the adaptability of VIA to infer cell-cycle stages based on label-free single-cell biophysical morphology (38 features, **see Supplementary Table S4 and Table S5**) profiled by our recently developed high-throughput imaging flow cytometer, called FACED[33]. VIA reliably reconstructed the continuous cell-cycle progressions from G1-S-G2/M phase of two different types of live breast cancer cells as validated by the single-cell fluorescent (DNA dye) images captured by the same system (**Methods**)(**Fig. 3i-k** for MCF7**, Supplementary Fig. S10** for MDA-MB231**)**. Intriguingly, according to the pseudotime ordered by VIA, not only can it reveal the known cell growth in size and mass[34], and general conservation of cell mass density[35] (as derived from the FACED images (**Methods**)) throughout the G1/S/G2 phases, but also a slow-down trend during the G1/S transition, consistent with the lower protein-accumulation rate during S phase[36] (**Fig. 3l, Supplementary Fig, S10 f,g**). The variation in biophysical textures (e.g. phase entropy) along the VIA pseudotime likely relates to known architectural changes of chromosomes and cytoskeletons during the cell cycles (**Fig. 3l, Fig. S10 f,g**). These results further substantiate the growing body of work[37,38,39,40] on imaging biophysical cytometry for gaining a mechanistic understanding of biological systems, especially when combined with omics analysis[41].

Concluding Remarks

Overall, VIA offers an advancement to TI methods to study a diverse range of single-cell omic data, including those targeted by many cell-atlas initiatives. By combining lazy-teleporting random walks and MCMC simulations, VIA relaxes common constraints on graph traversal and causality. This enables accurate lineage prediction that is robust to parameter configuration for a variety of complex topologies and rarer lineages obscured in other methods. Our stress tests showed that the modeled developmental landscape in other methods is vulnerable to user parameter choice which can incur fragmentation or spurious linkages, and consequently only yield biologically sensible lineages for a narrow sweet spot of parameters (See the summary in **Supplementary Fig. S16**). For example, due to algorithmic measures taken to restrict permissible graph-edge transitions and progressively reduce the inherent dimensionality (e.g. PCA followed by subsetting the number of diffusion components) other algorithms struggle to delineate obscure lineages and maintain neighborhood relationships. VIA's wider bandwidth of accuracy, superior runtime and preservation of global graph properties for very large datasets, offers a unique and well-suited approach for multifaceted exploratory analysis to uncover novel biological processes, potentially those deviated from healthy trajectories.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

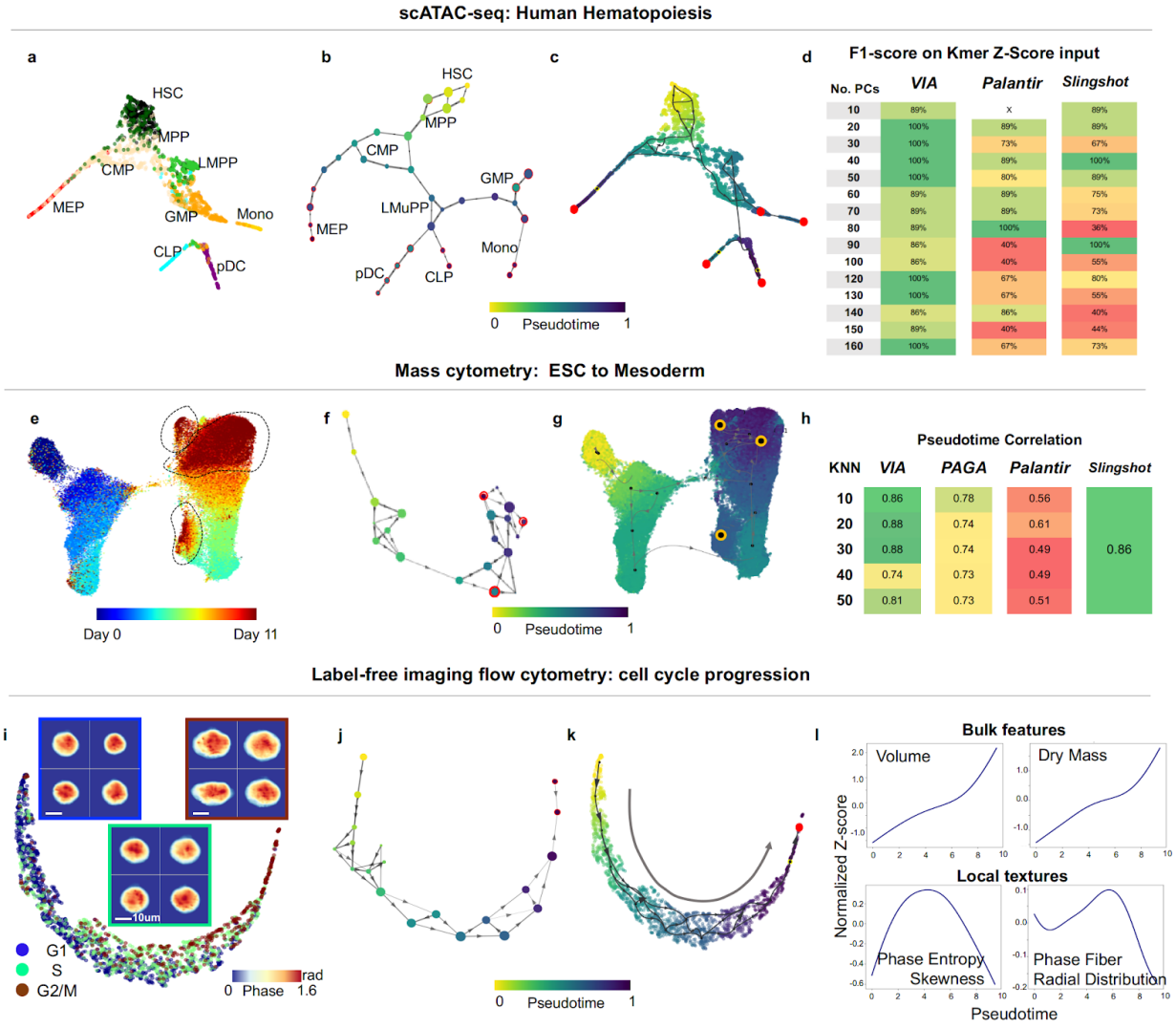


Figure 3 VIA infers trajectories in single-cell multi-omic and image datasets **(a)** Major lineages of human hematopoiesis (profiled by scATAC-seq) projected onto the UMAP embedding. Lineages are colored by FACS sorted labels[27]. **(b)** VIA cluster-graph topology colored by VIA pseudotime. **(c)** Trajectory, pseudotime and detected terminal states (red) projected onto the UMAP embedding. **(d)** F1-scores (on the k-mer Z-score input) for terminal state prediction by different TI methods (for a fixed KNN = 20). Terminal states include megakaryocyte–erythroid progenitor (MEP), common lymphoid progenitor (CLP), plasmacytoid dendritic cell (pDC) and monocytes (Mono) lineages. The comparisons show that VIA's accuracy remains high across a wide range of PCs. **(e)** Differentiation of mESC to mesoderm cells measured by single-cell mass cytometry. UMAP embedding is colored by different measurement time points (Day 0-11). **(f)** VIA cluster graph with 3 detected terminal nodes (red) and colored by pseudotime. **(g)** VIA results projected onto single-cell UMAP embedding shows 3 terminal states correspond to Day 10/11 regions. **(h)** Correlation of inferred pseudotime and day-labels achieved by different TI methods. The benchmark was done across different numbers of KNN (using all 28 antibodies). **(i)** Label-free cell cycle progression tracking based on FACED imaging cytometry. The PHATE embedding is constructed using 38 biophysical/morphological features computed from images of human breast cancer cells (MCF7) (See **Supplementary Fig. S10** for additional results using another breast cancer cell type (MDA-MB231)). The embedding is colored by the known cell cycle stages given by the DNA fluorescence images (obtained from the same system). **(j)** VIA graph topology colored by pseudotime. **(k)** VIA trajectory and pseudotime projected on embedding. **(l)** "Biophysical" feature expressions (Z-score normalized) over pseudotime. (See **Supplementary Table S4-5** for detailed definitions of the features).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

VIA Algorithm

VIA applies a scalable probabilistic method to infer cell state dynamics and differentiation hierarchies by organizing cells into trajectories along a pseudotime axis in a nearest-neighbor graph which is the basis for subsequent random walks. Single cells are represented by graph nodes that are connected based on their feature similarity, e.g. gene expression, transcription factor accessibility motif, protein expression or morphological features of cell images. A typical routine in VIA mainly consists of four steps:

1. **Accelerated and scalable cluster-graph construction**. VIA first represents the single-cell data in a k-nearest-neighbor (KNN) graph where each node is a cluster of single cells. The clusters are computed by our recently developed clustering algorithm, PARC[11]. In brief, PARC is built on hierarchical navigable small world (HNSW[58]) accelerated KNN graph construction and a fast community-detection algorithm (Leiden method[42]), which is further refined by data-driven pruning. The combination of these steps enables PARC to outperform other clustering algorithms in computational run-time, scalability in data size and dimension (without relying on subsampling of large-scale, high-dimensional single-cell data (>1 million cells)), and sensitivity of rare-cell detection. We employ the cluster-level topology, instead of a single-cell-level graph, for TI as it provides a coarser but clearer view of the key linkages and pathways of the underlying cell dynamics without imposing constraints on the graph edges. Together with the strength of PARC in clustering scalability and sensitivity, this step critically allows VIA to faithfully reveal complex topologies namely cyclic, disconnected and multifurcating trajectories (**Supplementary Fig. S1**).

2. **Probabilistic pseudotime computation**. The trajectories are then modeled in VIA as (i) lazy-teleporting random walk paths along which the pseudotime is computed and further refined by (ii) MCMC simulations. The root is a single cell chosen by the user.These two sub-steps are detailed as follows:

   (i) *Lazy-teleporting random walk*: We first compute the pseudotime as the expected hitting time of a *lazy-teleporting* random walk on an undirected cluster-graph generated in Step 1. The lazy-teleporting nature of this random walk ensures that as the sample size grows, the expected hitting time of each node does not converge to the stationary probability given by local node properties, but instead continues to incorporate the wider global neighborhood information[12]. Here we highlight the derivation of the closed form expression of the hitting time of this modified random walk with a detailed derivation in **Supplementary Note 2**.

   The cluster graph constructed in VIA is mathematically defined as a weighted connected graph **G** (V, E, W) with a vertex set V of n vertices (or nodes), i.e. $V = \{v_1, \cdots, v_n\}$ and an edge set E, i.e. a set of ordered pairs of distinct nodes. W is an $n \times n$ weight matrix that describes a set of edge weights between node i and j, $w_{ij} \geq 0$ are assigned to the edges $\left(v_i, v_j\right)$. For an undirected graph, $w_{ij} = w_{ji}$ The $n \times n$ probability transition matrix, P, of a standard random walk on this graph **G** can be given by

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$P = D^{-1}W \qquad (1)$$

where D is the $n \times n$ degree matrix, which is a diagonal matrix of the weighted sum of the degree of each node, i.e. the matrix elements are expressed as

$$d_{ij} = \begin{cases} \sum_k w_{ik} & , i = j \\ 0 & , i \neq j \end{cases} \qquad (2)$$

where k are the neighbouring nodes connected to node i. Hence, d_{ii} (which can be reduced as d_i) is the degree of node i. We next consider a *lazy* random walk, defined as Z, with probability $(1-x)$ of being lazy (where $0 < x < 1$), i.e. staying at the same node, then

$$Z = xP + (1-x)I \qquad (3)$$

where I is the identity matrix. When teleportation occurs with a probability $(1-\alpha)$, the modified lazy-teleporting random walk Z' can be written as follows, where J is an $n \times n$ matrix of ones.

$$Z' = \alpha Z + (1-\alpha)\frac{1}{n}J \qquad (4)$$

Here we adapt the concept of personalized PageRank vector, originally used for recording (or *ranking*) personal preferences of a web-surfer toward particular website pages[43], to *rank* the importance of other nodes (clusters of cells) to a given node, depending on the similarities among nodes (related to P in the graph), and the lazy-teleporting random walk characteristics in the graph (set by probabilities of teleporting and being lazy). Based on this concept, one could model the likelihood to transit from one node (cluster of cells) to another, and thus construct the pseudotime based on the hitting time, which is a parameter describing the expected number of steps it takes for a random walk that starts at node i and visit node j for the first time. Consider the teleporting probability of $(1-\alpha)$ and a seed vector s specifying the initial probability distribution across the n nodes (such that $\sum_m s_m = 1$, where s_m is the probability of starting at node m) the personalized PageRank vector $pr_\alpha(s)$ (which is defined as a column vector) is the unique solution to[56]

$$pr_\alpha(s)^T = \alpha pr_\alpha(s)^T Z + (1-\alpha)s^T . \qquad (5)$$

Substituting Z (Eq. (3)) into Eq. (5), we can express the personalized PageRank vector $pr_\alpha(s)$ in terms of the inverse of the $\boldsymbol{\beta}$-normalized Laplacian, $R_{\beta,NL}$ of the modified random walk (**Supplementary Note 2), i.e.**

$$pr_\alpha(s)^T = \beta s^T D^{-0.5} R_{\beta,NL} D^{0.5} , \quad (6)$$

where $\beta = \frac{2(1-\alpha)}{(2-\alpha)}$, and $R_{\beta,NL} = \sum_{m=1} \frac{\Phi_m \Phi_m^T}{[\beta + 2x(1-\beta)\eta_m]}$. Φ_m and η_m are the m^{th} eigenvector and eigenvalue of the normalized Laplacian. In the expression of $R_{\boldsymbol{\beta},\mathbf{NL}}$, the $\boldsymbol{\beta}$ and $\boldsymbol{x}$ regulate the weight of contribution in each eigenvalue-eigenvector pair of the summation such that the first eigenvalue-eigenvector pair (corresponding to the stationary distribution and given by the local-node degree-properties) remains included in the overall expression, but does not overwhelm the global information provided by subsequent 'eigen-pairs'. Moreover, computation of $R_{\boldsymbol{\beta},\mathbf{NL}}$ is not limited to a subset of the first k eigenvectors (bypassing the need for the user to select a suitable threshold or subset of eigenvectors) since the dimensionality is not on the order of

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

number of cells, but equal to the number of clusters and hence all eigenvalue-eigenvector pairs can be incorporated without causing a bottleneck in runtime.

The expected hitting time from node q to node r is given by[44],

$$h_\alpha(q,r) = \frac{[pr_\alpha(e_r)^T](r)}{d_r} - \frac{[pr_\alpha(e_r)^T](q)}{d_q} \qquad (7)$$

where e_i is an indicator vector with 1 in the i^{th} entry and 0 elsewhere (i.e. $s_m = 1$ if $m = i$ and $s_m = 0$ if $m \neq i$). We can substitute Eq. (6) into Eq. (7), making use of the fact that $\frac{1}{d_r} = \left[D^{-1}e_r\right](r)$, and $D^{-0.5}R_{\beta,NL}\,D^{-0.5}$ is symmetric, to obtain a closed form expression of the hitting time in terms of $R_{\beta,NL}$

$$h_\alpha(q,r) = \beta(e_r - e_q)^T D^{-0.5} R_{\beta,NL}\, D^{-0.5} e_r \quad (8)$$

(ii) *MCMC simulation*: The hitting time metric computed in Step-1 is used to infer graph-directionality. Instead of pruning edges in the 'reverse' direction, edge-weights are biased based on the time difference between nodes using the logistic function with growth factor b =1.

$$f(t) = \frac{1}{1+e^{-b\,(t_1 - t_0)}}$$

We then recompute the pseudotimes on the forward biased graph: Since there is no closed form solution of hitting times on a *directed* graph, we perform MCMC simulations (parallely processed to enable fast simulations of 1000s of teleporting, lazy random walks starting at the root node of the cluster graph) and use the first quartile of the simulated pseudotime values for a respective node as the refined pseudotime for that node relative to the root. This refinement step ensures that the pseudotime is robust to the spurious links (or conversely, links that are too weakly weighted) that can distort calculations based purely on the closed form solution of hitting times (**Supplementary Fig. S9d**). By using this 2-step pseudotime computation, VIA mitigates the issues of convergence issues and spurious edge-weights, both of which are common in random-walk pseudotime computation on large and complex datasets[12].

3. **Automated terminal-state detection.** The algorithm then uses the refined directed and weighted graph (the edges are re-weighted using the refined pseudotimes) to predict which nodes represent the terminal states based on a consensus vote of pseudotime and multiple vertex connectivity properties, including out-degree (i.e. the number of edges directing out from the node), closeness ***C(q)***, and betweenness ***B(q)***.

$$C(q) = \frac{1}{\sum_{q \neq r} l(q,r)}$$

$$B(q) = \sum_{r \neq q \neq t} \frac{\sigma_{rt}(q)}{\sigma_{rt}}$$

$l(q,r)$ is the distance between node q and node r (i.e. the sum of edges in a shortest path connecting them). σ_{rt} is the total number of shortest paths from node r to node t. $\sigma_{rt}(q)$ is the number of these

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

paths passing through node *q*. The consensus vote is performed on nodes that score above (or below for out-degree) the median in terms of connectivity properties. We show on multiple simulated and real biological datasets that VIA more accurately predicts the terminal states across a range of input data dimensions and key algorithm parameters than other methods attempting the same **(Supplementary Fig. S16).**

4. **Automated trajectory reconstruction**. VIA then identifies the most likely path of each lineage by computing the likelihood of a node traversing towards a particular terminal state (e.g. differentiation). These lineage likelihoods are computed as the visitation frequency under lazy-teleporting MCMC simulations from the root to a particular terminal state, i.e. the probability of *node i* reaching *terminal-state j* as the number of times *cell i* is visited along a successful path (i.e. *terminal-state j* is reached) divided by the number of times *cell i* is visited along all of the simulations. In contrast to other trajectory reconstruction methods which compute the shortest paths between root and terminal node[1,2], the lazy-teleporting MCMC simulations in VIA offer a probabilistic view of pathways under relaxed conditions that are not only restricted to the random-walk along a tree-like graph, but can also be generalizable to other types of topologies, such as cyclic or connected/disconnected paths. In the same vein, we avoid confining the graph to an absorbing Markov chain[13,3] (AMC) as this places prematurely strict / potentially inaccurate constraints on node-to-node mobility and can impede sensitivity to cell fates (as demonstrated by VIA's superior cell fate detection across numerous datasets (**Supplementary Fig. S16**).

Downstream visualization and analysis

VIA generates a visualization that combines the network topology and single-cell level pseudotime/lineage probability properties onto an embedding based on UMAP or PHATE. Generalized additive models (GAMs) are used to draw edges found in the high-dimensional graph onto the lower dimensional visualization (**Fig. 1**). An unsupervised downstream analysis of cell features (e.g. marker gene expression, protein expression or image phenotype) along pseudotime for each lineage is performed (**Fig. 1**). Specifically, VIA plots the expression of features across pseudotime for each lineage by using the lineage likelihood properties to weight the GAMs. A cluster-level lineage pathway is automatically produced by VIA to visualize feature heat maps at the cluster-level along a lineage-path to see the regulation of genes. VIA provides the option of gene imputation before plotting the lineage specific gene trends. The imputation is fast as it relies on the single-cell KNN (scKNN) graph computed in Step 1. Using an affinity-based imputation method[45], this step computes a "diffused" transition matrix on the scKNN graph used to impute and denoise the original gene expressions.

Benchmarked Methods

The methods were mainly chosen based on their superior performance in a recent large-scale benchmarking study[4], including a select few recent methods claiming to supersede those in the study. Specifically, recent and popular methods exhibiting reasonable scalability, and automated cell fate prediction in multi-lineage trajectories were favoured as candidates for benchmarking (See **Supplementary Table S1** for the key characteristics of methods). Performance stress-tests in terms of lineage detection of each biological dataset, and pseudotime correlation for time-series data were

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

conducted over a range of key input parameters (e.g. numbers of k-nearest neighbors, highly variable genes (HVGs), and principal components (PCs)) and pre-processing protocols (see **Fig. 2m,p, Supplementary Fig. 16**). All comparisons were run on a computer with an Intel(R) Xeon (R) W-2123 central processing unit (3.60GHz, 8 cores) and 126 GB RAM.

Quantifying terminal state prediction accuracy for parameter tests was done using the F1-score, defined as the harmonic mean of recall and precision and calculated as:

$$F_1 = \frac{tp}{tp + 0.5(fp+fn)}$$

Where *tp* is a true-positive: the identification of a terminal cluster that is in fact a final differentiated cell fate; *fp* is a false positive identification of a cluster as terminal when in fact it represents an intermediate state; and *fn* is a false negative where a known cell fate fails to be identified

PAGA[28]. It uses a cluster-graph representation to capture the underlying topology. PAGA computes a unified pseudotime by averaging the single-cell level diffusion pseudotime computed by DPT, but requires manual specification of terminal cell fates and clusters that contribute to lineages of interest in order to compare gene expression trends across lineages.

Palantir[2]. It uses diffusion-map[46.] components to represent the underlying trajectory. Pseudotimes are computed as the shortest path along a KNN-graph constructed in a low-dimensional diffusion component space, with edges weighted such that the distance between nodes corresponds to the diffusion pseudotime[47.] (DPT). Terminal states are identified as extrema of the diffusion maps that are also outliers of the stationary distribution. The lineage-likelihood probabilities are computed using Absorbing Markov Chains (constructed by removing outgoing edges of terminal states, and thresholding reverse edges).

Slingshot[1]. It is designed to process low-dimensional embeddings of the single-cell data. By default Slingshot runs clustering based on Gaussian mixture modeling and recommends using the first few PCs as input. Slingshot connects the clusters using a minimum spanning tree and then fits principle curves for each detected branch. It uses the orthogonal projection against each principal curve to fit a separate pseudotime for each lineage, and hence the gene expressions cannot be compared across lineages. Also, the runtimes are prohibitively long for large datasets or high input dimensions.

CellRank[13]. This method combines the information of RNA velocity (computed using scVelo[48.]) and gene-expression to infer trajectories. Given it is mainly suited for the scRNA-seq data, with the RNA-velocity computation limiting the overall runtime for larger dataset, we limit our comparison to the pancreatic dataset which the authors of CellRank used to highlight its performance.

Simulated Data

We employed the DynToy[4] (https://github.com/dynverse/dyntoy) package, which generates synthetic single-cell gene expression data (~1000 cells x 1000 'genes'), to simulate different complex trajectory models. Using these datasets, we tested that VIA consistently and more accurately captures both tree and non-tree like structures (multifurcating, cyclic, and disconnected) compared to other methods

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(Supplementary Fig. S1). All methods are subject to the same data pre-processing steps, PCA dimension reduction and root-cell to initialize the path.

Multifurcating structure. This dataset consists of 1000 'cells' multifurcating into 4 terminal states. VIA robustly captures all four terminal cell fates across a range of input PCs and the pseudotimes are well inferred relative to the root node **(Supplementary Fig. S1a)**. Note that two terminal states (M2 and M8), which are very close to each other, are easily merged by the other methods (Slingshot, Palantir and PAGA).

Cyclic structure. We ran VIA and other methods for different values of K nearest neighbors. VIA unambiguously shows a cyclic network for a range of K (in KNN). Slingshot does not use a KNN parameter and shows 3 fragmented different lineages (top to bottom). PAGA fails to capture the connected cyclic structure at K = 10 and 5, while Palantir visually shows a linear (K = 10, 30) or disconnected structure (K = 5). Van den Berge et al[57] note that the challenge of cyclic trajectory reconstruction is also common in other popular methods, such as Monocle3 that consistently fragments or fits branching structures onto cyclic simulated datasets.

Disconnected structure. This dataset comprises two disconnected trajectories (T1 and T2). T1 is cyclic with an extra branch (M5 to M6), T2 has a bifurcation at M3 (**Supplementary Fig. S1c)**. VIA captures the two disconnected structures as well as the M6 branch in the cyclic structure, and the bifurcation in the smaller structure. PAGA captures the underlying structure at PC = 20 but becomes fragmented for other numbers of PCs. Palantir also yields multiple fragments and is not able to capture the overall structure, while Slingshot (using the default clustering based on Gaussian mixture modeling) connects T1 and T2, and only captures one of the bifurcations in T1.

Biological Data

The pre-processing steps described below for each dataset are not included in the reported runtimes as these steps are typically very fast, (typically less than 1-10% of the total runtime depending on the method. E.g. only a few minutes for pre-processing 100,000s of cells) and only need to be performed once as they remain the same for all subsequent analyses. It should also be noted that visualization (e.g. UMAP, t-SNE) are not included in the runtimes. VIA provides a subsampling option at the visualization stage to accelerate this process for large datasets without impacting the previous computational steps. However, to ensure fair comparisons between TI methods (e.g. other methods do not have an option to compute the embedding on a subsampled input and transfer the results between the full trajectory and the sampled visualization, or rely on a slow version of tSNE), we simply provide each TI method with a pre-computed visualization embedding on which the computed results are projected.

ScRNA-seq of mouse pre-B cells. This dataset[26] models the pre-BI cell (Hardy fraction C') process during which cells progress to the pre-BII stage and B cell progenitors undergo growth arrest and differentiation. Measurements were obtained at 0, 2, 6, 12, 18 and 24 hours (h) for a total of 313 cells x 9,075 genes. We follow a standard Scanpy preprocessing recipe[49] that filters cells with low counts, and genes that occur in less than 3 cells. The filtered cells are normalized by library size and log transformed. The top 5000 highly variable genes (HVG) are retained. Cells are renormalized by library count and scaled to unit variance and zero mean. VIA identifies the terminal state at 18-24 h and accurately

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

recapitulates the gene expression trends[26] along inferred pseudotime of *IgII1*, *Slc7a5*, *Fox01*, *Myc*, *Ldha* and *Lig4*. (**Supplementary Fig. S2a).** We show the results generalize across a range of PCs for two values of K of the graph with higher accuracy in locating the later cell fates than Slingshot and Palantir. (**Supplementary Fig. S2b).**

ScRNA-seq of human CD34+ bone marrow cells. This is a scRNA-seq dataset of 5800 cells representing human hematopoiesis[2]. We used the filtered, normalized and log-transformed count matrix provided by Setty et al[2]., with PCA performed on all the remaining genes. The cells were annotated using SingleR[50] which automatically labeled cells based on the hematopoietic reference dataset Novershtern Hematopoietic Cell Data - GSE24759[51]. The annotations are in agreement with the labels inferred by Setty et al. for the 7 clusters, including the root HSCs cluster that differentiates into 6 different lineages: monocytes, erythrocytes, and B cells, as well as the less populous megakaryocytes, cDCs and pDCs. VIA consistently identifies these lineages across a wider range of input parameters and data dimensions (e.g. the number of K and PCs provided as input to the algorithms see **Fig. 2p,** and **Supplementary Fig. S3c**). Notably, the upregulated gene expression trends of the small populations can be recovered in VIA, i.e. pDC and cDC show elevated CD123 and CSF1R levels relative to other lineages, and the upregulated CD41 expression in megakaryocytes (**Supplementary Fig. S3-S4)**.

ScRNA-seq of human embryoid body. This is a midsized scRNA-seq dataset of 16,825 human cells in embryoid bodies (EBs)[15]. We followed the same pre-processing steps as Moon et al. to filter out dead cells and those with too high or low library count. Cells are normalized by library count followed by square root transform. Finally the transformed counts are scaled to unit variance and zero mean. The filtered data contained 16825 cells × 17580 genes. PCA is performed on the processed data before running each TI method. VIA identifies 6 cell fates, which, based on the upregulation of marker genes as cells proceed towards respective lineages, are in accord with the annotations given by Moon et al., (See the gene heatmap and changes in gene expression along respective lineage trajectories in **Supplementary Fig. S5).** Note that Palantir and Slingshot do not capture the cardiac cell fate, and Slingshot also misses the neural crest **(**see the F1-scores summary for terminal state detection **Supplementary Fig. S5).**

ScRNA-seq of mouse organogenesis cell atlas. This is a large and complex scRNA-seq dataset of mouse organogenesis cell atlas (MOCA) consisting of 1.3 million cells[6]. The dataset contains cells from 61 embryos spanning 5 developmental stages from early organogenesis (E9.5-E10.5) to organogenesis (E13.5). Of the 2 million cells profiled, 1.3 million are 'high-quality' cells that are analysed by VIA. The runtime is approximately 40 minutes which is in stark contrast to the next fastest tool Palantir which takes 4 hours (excluding visualization). The authors of MOCA manually annotated 38 cell-types based on the differentially expressed genes of the clusters. In general, each cell type exclusively falls under one of 10 major and disjoint trajectories inferred by applying Monocle3 to the UMAP of MOCA. The authors attributed the disconnected nature of the 10 trajectories to the paucity of earlier stage common predecessor cells. We followed the same steps as Cao et al.[6] to retain high-quality cells (i.e. remove cells with less than 400 mRNA, and remove doublet cells and cells from doubled derived sub-clusters). PCA was applied to the top 2000 HVGs with the top 30 PCs selected for analysis. VIA analyzed the data in the high-dimensional PC space. We bypass the step in Monocle3[6] which applies UMAP on the PCs prior to TI as this incurs an additional bias from choice of manifold-learning parameters and a further loss in

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

neighborhood information. As a result, VIA produces a more connected structure with linkages between some of the major cell types that become segregated in UMAP (and hence Monocle3), and favors a biologically relevant interpretation (**Fig. 2, Supplementary Fig. S11**). A detailed explanation of these connections (graph-edges) extending between certain major groups using references to literature on organogenesis is presented in **Supplementary Note 3.**

ScRNA-seq of murine endocrine development[5]. This is an scRNA-seq dataset of E15.5 murine pancreatic cells spanning all developmental stages from an initial endocrine progenitor-precursor (EP) state (low level of *Ngn3*, or *Ngn3*^{low}), to the intermediate EP (high level of *Ngn3*, or *Ngn3*^{high}) and Fev^{+} states, to the terminal states of hormone-producing alpha, beta, epsilon and delta cells5.5. Following steps by Lange et al[13], we preprocessed the data using scVelo to filter genes, normalize each cell by total counts over all genes, keep the top most variable genes, and take the log-transform. PCA was applied to the processed gene matrix. We assessed the performance of VIA and other TI methods (CellRank, Palantir, Slingshot) across a range of number of retained HVGs and input PCs (**Fig. 2m**, **Supplementary Fig. S6)**.

ScATAC-seq of human bone marrow cells. This scATAC-seq data profiles 3072 cells isolated from human bone marrow using fluorescence activated cell sorting (FACS), yielding 9 populations[27]: HSC, MPP, CMP, CLP, LMPP, GMP, MEP, mono and plasmacytoid DCs (**Fig. 3a and Supplementary Fig. S7**). We examined TI results for two different preprocessing pipelines to gauge how robust VIA is on the scATAC-seq analysis which is known to be challenging for its extreme intrinsic sparsity. We used the pre-processed data consisting of PCA applied to the z-scores of the transcription factor (TF) motifs used by Buenrostro et a[27]. Their approach corrects for batch effects in select populations and weighting of PCs based on reference populations and hence involves manual curation. We also employed a more general approach used by Chen et al.[31] which employs ChromVAR to compute k-mer accessibility z-scores across cells. VIA infers the correct trajectories and the terminal cell fates for both of these inputs, again across a wide range of input parameters (**Fig. 3d and Supplementary Fig. S7**).

ScRNA-seq and scATAC-seq of *Isl1*+ cardiac progenitor cells. This time-series dataset captures murine *Isl1*+ cardiac progenitor cells (CPCs) from E7.5 to E9.5 characterized by scRNA-seq (197 cells) and scATAC-seq (695 cells)[20]. The *Isl1*+ CPCs are known to undergo multipotent differentiation to cardiomyocytes or endothelial cells. For the scRNA-seq data, the quality filtered genes and the size-factor normalized expression values are provided by Jia et al.[20] as a "Single Cell Expression Set" object in R. Similarly, the cells in the scATAC-seq experiment were provided in a "SingleCellExperiment" object with low quality cells excluded from further analysis. The accessibility of peaks was transformed to a binary representation as input for TF-IDF (term frequency-inverse document frequency) weighting prior to singular value decomposition (SVD). The highlighted TF motifs in the heatmap (**Fig. 2j**) correspond to those highlighted by Jia et al. We tested the performance when varying the number of SVDs used. We also considered the outcome when merging the scATAC-seq and scRNA-seq data using Seurat3[52]. Despite the relatively low cell count of both datasets, and the relatively under-represented scRNA-seq cell count, the two datasets overlapped reasonably well and allowed us to infer the expected lineages in an unsupervised manner (**Fig. 2d and Supplementary Fig. S8**. In contrast, Jia et al., performed a supervised TI by manually selecting cells relevant to the different lineages (for the scATAC-seq cells) and choosing the two diffusion components that best characterize the developmental trajectories in low dimension[20].

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Mass cytometry data of mouse embryonic stem cells (mESC). This is a mass cytometry (or CyTOF) dataset, consisting of 90,000 cells and 28 antibodies (corresponding to ~7000 cells each from Day 0-11 measurements), that represents differentiation of mESC to mesoderm cells[32]. An arcsinh transform with a scaling factor of 5 was applied on all features - a standard procedure for CyTOF datasets, followed by normalization to unit variance and zero mean. Given the small feature set, no PCA is required **(Supplementary Fig. S9)**. VIA identifies 3 main terminal states corresponding to Day 11 and Day 10, Palantir on the other hand identifies three Terminal states that all correspond to Days in the first half of the experiment and the pseudotime is heavily influenced by the root node being very weakly connected to the other stages of the process. Slingshot appears to capture the overall pseudotime but the 6 lineages imposed onto the low dimensional representation are difficult to interpret and distinguish. To improve Palantir performance we used 5000 waypoints but this takes almost 20 minutes to complete (excluding time taken for embedding the visualization). VIA runs in ~3 minutes and produces results consistent with the known ordering. The pseudotime reflects the range of Days very well, even capturing the small population of Day 11 cells on the left hand side of the Day 6 cells in the embedding **(Fig. 3, and Supplementary Fig. S9)**.

Single-cell biophysical phenotypes derived from imaging flow cytometry. This is the in-house dataset of single-cell biophysical phenotypes of two different human breast cancer types (MDA-MB231 and MCF7). Following our recent image-based biophysical phenotyping strategy[53,54], we defined the spatially-resolved biophysical features of a cell in a hierarchical manner based on both bright-field and quantitative phase images captured by the FACED imaging flow cytometer (i.e., from the bulk features to the subcellular textures). At the bulk level, we extracted the cell size, dry mass density, and cell shape. At the subcellular texture level, we parameterized the global and local textural characteristics of optical density and mass density at both the coarse and fine scales (e.g., local variation of mass density, its higher-order statistics, phase entropy radial distribution etc.). This hierarchical phenotyping approach[53,54] allowed us to establish a single-cell biophysical profile of 38 features, which were normalized based on the z-score (**See Supplementary Table S4 and Table S5**). All these features, without any PCA, are used as input to VIA. In order to weigh the features, we use a mutual information classifier to rank the features, based on the integrated fluorescence intensity of the fluorescence FACED images of the cells (which serve as the ground truth of the cell-cycle stages). Following normalization, the top 3 features (which relate to cell size) are weighted (using a factor between 3-10).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Imaging flow cytometry experiment

FACED imaging flow cytometer setup

A multimodal FACED imaging flow cytometry (IFC) platform was used to obtain the quantitative phase and fluorescence images of single cells in microfluidic flow at an imaging throughput of ~70,000 cells/sec. The light source consisted of an Nd:YVO picosecond laser (center wavelength = 1064 nm, Time-Bandwidth) and a periodically-poled lithium niobate (PPLN) crystal (Covesion) for second harmonic generation of a green pulsed beam (center wavelength = 532 nm) with a repetition rate of 20 MHz. The beam was then directed to the FACED module, which mainly consists of a pair of almost-parallel plane mirrors. This module generated a linear array of 50 beamlets (foci) which were projected by an objective lens (40X, 0.6NA, MRH08430, Nikon) on the flowing cells in the microfluidic channel for imaging. Each beamlet was designed to have a time delay of 1 ns with the neighboring beamlet in order to minimize the fluorescence crosstalk due to the fluorescence decay. Detailed configuration of the FACED module can be referred to Wu et al.[33]. The epi-fluorescence image signal was collected by the same objective lens and directed through a band-pass dichroic beamsplitter (center: 575nm, bandwidth: 15nm). The filtered orange fluorescence signal was collected by the photomultiplier tube (PMT) (rise time: 0.57 ns, Hamamatsu). On the other hand, the transmitted light through the cell was collected by another objective lens (40X, 0.8NA, MRD07420, Nikon). The light was then split equally by the 50:50 beamsplitter into two paths, each of which encodes different phase-gradient image contrasts of the same cell (a concept similar to Scherlien photography[55]). The two beams are combined, time-interleaved, and directed to the photodetector (PD) (bandwidth: >10 GHz, Alphalas) for detection. The signals obtained from both PMT and PD were then passed to a real-time high-bandwidth digitizer (20 GHz, 80 GS/s, Lecroy) for data recording.

Cell culture and preparation

MDA-MB231 (ATCC) and MCF7 (ATCC), which are two different breast cancer cell lines, were used for the cell cycle study. The culture medium for MDA-MB231was ATCC modified RPMI 1640 (Gibco) supplemented with 10% fetal bovine serum (FBS) (Gibco) and 1% antibiotic-antimycotic (Anti-Anti) (Gibco), while that for MCF7 was DMEM supplemented with 10% FBS (Gibco) and 1% Anti-Anti (Gibco). The cells were cultured inside an incubator under 5% CO_2 and 37°C, and subcultured twice a week. 1e6 cells were pipetted out from each cell line and stained with Vybrant DyeCycle orange stain (Invitrogen).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Data Availability

Data used in Figures 1-3 as well as Supplementary Figures S1-S15) is available on:

1. Pancreatic data: Gene Expression Omnibus (GEO) under accession code GSE132188.
2. Cardiac progenitor data is available from the ENA repository under the accession code PRJEB23303 or from [https://github.com/loosolab/cardiac-progenitors].
3. B-cell: STATegraData GitHub repository. [https://github.com/STATegraData/STATegraData]
4. Mass cytometry mesoderm: Cytobank [https://community.cytobank.org/cytobank/experiments/71953].
5. Raw and processed data for scRNA-seq Human Hematopoeisis are available through the Human Cell Atlas data portal at https://data.humancellatlas.org/explore/projects/091cf39b-01bc-42e5-9437-f419a66c8a45.
6. Embryoid Body: Mendeley Data repository at https://doi.org/10.17632/v6n743h5ng.1.
7. Mouse Organogenesis : NCBI Gene Expression Omnibus under accession number GSE119945
8. FACED cell cycle: https://github.com/ShobiStassen/VIA and on FigShare https://doi.org/10.6084/m9.figshare.13601405.v1
9. scATAC-seq Hematopoiesis: GEO: GSE96772. Processed scATAC-seq data, which include PC values and TF scores per cell can be found in Data S1. of https://doi.org/10.1016/j.cell.2018.03.074
10. Toy Data: https://github.com/ShobiStassen/VIA

Code Availability

VIA is available as a pip installable python library "pyVIA" with tutorials and sample data available on https://github.com/ShobiStassen/VIA and https://pypi.org/project/pyVIA/

References

1. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 19, 477 (2018).
2. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir [published correction appears in Nat Biotechnol. 2019 Oct;37(10):1237]. Nat Biotechnol. 2019;37(4):451-460. doi:10.1038/s41587-019-0068-4
3. Qiu, X., Mao, Q., Tang, Y. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 14, 979–982 (2017). https://doi.org/10.1038/nmeth.4402
4. Saelens, W., Cannoodt, R., Todorov, H. et al. A comparison of single-cell trajectory inference methods. Nat Biotechnol 37, 547–554 (2019). https://doi.org/10.1038/s41587-019-0071-9
5. Bastidas-Ponce, A. et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinegenesis. Development 146, (2019).
6. Cao, J. et al. Comprehensive single- cell transcriptional profiling of a multicellular organism. Science 357,661–667 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

7. Packer, J. S. et al. A lineage- resolved molecular atlas of C. elegans embryogenesis at single- cell resolution.Science 365, eaax1971 (2019).
8. Cao, J., Spielmann, M., Qiu, X. et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502 (2019).
9. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single- cell resolution.Science 360, eaar5780 (2018).
10. Litviňuková, M., Talavera-López, C., Maatz, H. et al. Cells of the adult human heart. Nature (2020).
11. Stassen SV, Siu DMD, Lee KCM, Ho JWK, So HKH, Tsia KK. PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells. Bioinformatics. 2020 May 1;36(9):2778-2786. doi: 10.1093/bioinformatics/btaa042.
12. Ulrike von Luxburg, Agnes Rad, Matthias Hein. Hitting and Commute Times in Large Random Neighborhood Graphs. Journal of Machine Learning Research 15, 1751-1798 (2014)
13. Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, Fabian J. Theis. CellRank for directed single-cell fate mapping. bioRxiv 2020.10.19.345983; doi: https://doi.org/10.1101/2020.10.19.345983
14. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Software.* 3, 861 (2018).
15. Moon, K.R., van Dijk, D., Wang, Z. et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 37, 1482–1492 (2019). https://doi.org/10.1038/s41587-019-0336-3
16. Tam PP, Behringer RR. Mouse gastrulation: the formation of a mammalian body plan. Mech Dev. 1997;68(1-2):3-25. doi:10.1016/s0925-4773(97)00123-8
17. Chin AM, Hill DR, Aurora M, Spence JR. Morphogenesis and maturation of the embryonic and postnatal intestine. Semin Cell Dev Biol. 2017 Jun;66:81-93. doi: 10.1016/j.semcdb.2017.01.011. Epub 2017 Feb 1.
18. Gilbert SF. Developmental Biology. 6th edition. Sunderland (MA): Sinauer Associates; 2000. The Neural Crest. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10065/
19. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program, Nature (2019) https://doi.org/10.1038/s41586-019-1629-x
20. Jia G, Preussner J, Chen X, Guenther S, Yuan X, Yekelchyk M, Kuenne C, Looso M, Zhou Y, Teichmann S, Braun T. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. Nat Commun. 2018 Nov 19;9(1):4877.
21. Tanya E. Foley, Bradley Hess, Joanne G. A. Savory, Randy Ringuette, David Lohnes.Role of Cdx factors in early mesodermal fate decisions.Development 2019 146: dev170498 doi: 10.1242/dev.170498 Published 1 April 2019
22. Yao Y, Yao J, Boström KI. SOX Transcription Factors in Endothelial Differentiation and Endothelial-Mesenchymal Transitions. Front Cardiovasc Med. 2019;6:30. Published 2019 Mar 28. doi:10.3389/fcvm.2019.00030
23. Potta SP, Liang H, Winkler J, Doss MX, Chen S, Wagh V, Pfannkuche K, Hescheler J, Sachinidis A. Isolation and functional characterization of alpha-smooth muscle actin expressing

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

712 cardiomyocytes from embryonic stem cells. Cell Physiol Biochem. 2010;25(6):595-604. doi:
713 10.1159/000315078. Epub 2010 May 18. PMID: 20511704.
714 24. Warkman AS, Whitman SA, Miller MK, Garriock RJ, Schwach CM, Gregorio CC, Krieg PA.
715 Developmental expression and cardiac transcriptional regulation of Myh7b, a third myosin heavy
716 chain in the vertebrate heart. Cytoskeleton (Hoboken). 2012 May;69(5):324-35. doi:
717 10.1002/cm.21029. Epub 2012 Apr 30. Erratum in: Cytoskeleton (Hoboken). 2012
718 Dec;69(12):1086. PMID: 22422726; PMCID: PMC4734749.
719 25. Mahmoud AI, Kocabas F, Muralidhar SA, et al. Meis1 regulates postnatal cardiomyocyte cell
720 cycle arrest. Nature. 2013;497(7448):249-253. doi:10.1038/nature12054
721 26. Gomez-Cabrero, D., Tarazona, S., Ferreirós-Vidal, I. et al. STATegra, a comprehensive
722 multi-omics dataset of B-cell differentiation in mouse. Sci Data 6, 256 (2019).
723 https://doi.org/10.1038/s41597-019-0202-7
724 27. Jason D. Buenrostro, M. Ryan Corces, Caleb A. Lareau, Beijing Wu, Alicia N. Schep, Martin J.
725 Aryee, Ravindra Majeti, Howard Y. Chang, William J. Greenleaf, Integrated Single-Cell Analysis
726 Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation, Cell, 173,
727 1535-1548.e16, (2018) https://doi.org/10.1016/j.cell.2018.03.074.
728 28. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through
729 a topology preserving map of single cells. Genome Biol. 20, 59 (2019).
730 29. Gutierrez GD, Gromada J, Sussel L. Heterogeneity of the Pancreatic Beta Cell. Front Genet.
731 2017;8:22. Published 2017 Mar 6. doi:10.3389/fgene.2017.00022
732 30. Krentz NAJ, Lee MYY, Xu EE, Sproul SLJ, Maslova A, Sasaki S, Lynn FC. Single-Cell
733 Transcriptome Profiling of Mouse and hESC-Derived Pancreatic Progenitors. Stem Cell Reports.
734 2018 Dec 11;11(6):1551-1564. doi: 10.1016/j.stemcr.2018.11.008. PMID: 30540962; PMCID:
735 PMC6294286.
736 31. Chen, H., Lareau, C., Andreani, T. *et al.* Assessment of computational methods for the analysis of
737 single-cell ATAC-seq data. *Genome Biol* 20, 241 (2019).
738 https://doi.org/10.1186/s13059-019-1854-5
739 32. Ko, M.E., Williams, C.M., Fread, K.I. *et al.* FLOW-MAP: a graph-based, force-directed layout
740 algorithm for trajectory mapping in single-cell time course datasets. *Nat Protoc* 15, 398–420
741 (2020). https://doi.org/10.1038/s41596-019-0246-3
742 33. Wu J. L., Xu Y. Q., Xu J. J., Wei X. X., Chan A. C. S., Tang A. H. L., Lau A. K. S., Chung B. M.
743 F., Cheung Shum H., Lam E. Y., Wong K. K. Y., Tsia K. K., "Ultrafast laser-scanning time-stretch
744 imaging at visible wavelengths," Light Sci. Appl. 6(1), e16196 (2016).10.1038/lsa.2016.196
745 34. Popescu G, Park Y, Lue N, Best-Popescu C, Deflores L, Dasari RR, Feld MS, Badizadegan K.
746 Optical imaging of cell mass and growth dynamics. Am J Physiol Cell Physiol. 2008
747 Aug;295(2):C538-44. doi: 10.1152/ajpcell.00121.2008. Epub 2008 Jun 18.
748 35. Kyoohyun Kim, Jochen Guck The Relative Densities of Cytoplasm and Nuclear Compartments
749 Are Robust against Strong PerturbationBiophysical Journal. Volume 119, Issue 10, 17 November
750 2020, Pages 1946-1957
751 36. Kafri R, Levy J, Ginzberg MB, Oh S, Lahav G, Kirschner MW. Dynamics extracted from fixed
752 cells reveal feedback linking cell growth to cell cycle. Nature. 2013 Feb 28;494(7438):480-3. doi:
753 10.1038/nature11897. PMID: 23446419; PMCID: PMC3730528.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

37. Park SR, Namkoong S, Friesen L, Cho CS, Zhang ZZ, Chen YC, Yoon E, Kim CH, Kwak H, Kang HM, Lee JH. Single-Cell Transcriptome Analysis of Colon Cancer Cell Response to 5-Fluorouracil-Induced DNA Damage. Cell Rep. 2020 Aug 25;32(8):108077. doi: 10.1016/j.celrep.2020.108077.
38. Zangle TA, Teitell MA. Live-cell mass profiling: an emerging approach in quantitative biophysics. Nat Methods. 2014 Dec;11(12):1221-8. doi: 10.1038/nmeth.3175. PMID: 25423019; PMCID: PMC4319180.
39. Tse HT, Gossett DR, Moon YS, Masaeli M, Sohsman M, Ying Y, Mislick K, Adams RP, Rao J, Di Carlo D. Quantitative diagnosis of malignant pleural effusions by single-cell mechanophenotyping. Sci Transl Med. 2013 Nov 20;5(212):212ra163. doi: 10.1126/scitranslmed.3006559. PMID: 24259051.
40. Otto, O., Rosendahl, P., Mietke, A. *et al.* Real-time deformability cytometry: on-the-fly cell mechanical phenotyping. *Nat Methods* 12, 199–202 (2015). https://doi.org/10.1038/nmeth.3281
41. Kimmerling, R.J., Prakadan, S.M., Gupta, A.J. *et al.* Linking single-cell measurements of mass, growth rate, and gene expression. *Genome Biol* 19, 207 (2018). https://doi.org/10.1186/s13059-018-1576-0
42. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z
43. Langville, Amy N., and Carl D. Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.
44. Chung F., Zhao W. (2010) PageRank and Random Walks on Graphs. In: Katona G.O.H., Schrijver A., Szőnyi T., Sági G. (eds) Fete of Combinatorics and Computer Science. Bolyai Society Mathematical Studies, vol 20. Springer, Berlin, Heidelberg.
45. van Dijk D, Sharma R, Nainys J, et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell. 2018;174(3):716-729.e27. doi:10.1016/j.cell.2018.05.061
46. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl Acad. Sci. USA 102,7426–7431 (2005).
47. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016;13(10):845-848. doi:10.1038/nmeth.3971
48. Bergen, V., Lange, M., Peidli, S. et al. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 38, 1408–1414 (2020). https://doi.org/10.1038/s41587-020-0591-3
49. Zheng GX, Terry JM, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Jan 16;8:14049. doi: 10.1038/ncomms14049.
50. Aran D et al., (2019). "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage." Nat. Immunol., 20, 163-172
51. Novershtern N. et al., Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011 Jan 21;144(2):296-309.
52. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6.

bioRxiv preprint doi: https://doi.org/10.1101/2021.02.10.430705; this version posted February 11, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

53. Siu, KCM Lee, MCK Lo, SV Stassen, M Wang, IZQ Zhang, HKH So. Deep-learning-assisted biophysical imaging cytometry at massive throughput delineates cell population heterogeneity. Lab on a Chip 20 (20), 3696-3708
54. KCM Lee, M Wang, KSE Cheah, GCF Chan, HKH So, KKY Wong, KK Tsia.. Quantitative phase imaging flow cytometry for ultra-large-scale single-cell biophysical phenotyping. Cytometry Part A 95 (5), 510-520
55. Wenwei Yan Jianglai Wu Kenneth K. Y. Wong Kevin K. Tsia, A high-throughput all-optical laser-scanning imaging flow cytometer with biomolecular specificity and subcellular resolution, J. Biophotonics (2017) https://onlinelibrary.wiley.com/doi/abs/10.1002/jbio.201700178
56. F. Chung and S.-T. Yau, Discrete Green's Functions. Journal of Combinatorial Theory, Series A, 91(1-2) (2000), pp. 191–214
57. Van den Berge, K., Roux de Bézieux, H., Street, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Communications
58. Yury A. Malkov, D. Yashunin. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graph, Computer Science, Medicine, Mathematics, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020