PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes

Zheng Hu (zheng.hu@siat.ac.cn)
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences https://orcid.org/0000-0003-1552-0060

Kun Wang
Xiamen University https://orcid.org/0009-0009-6503-4678

Liangzhen Hou
University of Macau

Xin Wang
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences https://orcid.org/0000-0001-7440-1075

Xiangwei Zhai
Sun Yat-Sen University https://orcid.org/0000-0003-2967-5573

Zhaolian Lu
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Zhike Zi
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences https://orcid.org/0000-0002-7601-915X

Weiwei Zhai
Institute of Zoology, Chinese Academy of Sciences https://orcid.org/0000-0001-7938-0226

Xionglei He
Sun Yat-Sen University https://orcid.org/0000-0003-1050-802X

Christina Curtis
Stanford University School of Medicine https://orcid.org/0000-0003-0166-3802

Da Zhou
Xiamen University https://orcid.org/0000-0002-0272-6644

Article

Keywords:

Posted Date: August 1st, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2197712/v1

Additional Declarations: Yes there is potential Competing Interest. C.C. is an advisor and stockholder in Grail, Ravel, and DeepCell and an advisor to Genentech, Bristol Myers Squibb, 3T Biosciences, and NanoString. Other authors declare no competing interests.

Version of Record: A version of this preprint was published at Nature Biotechnology on July 31st, 2023. See the published version at https://doi.org/10.1038/s41587-023-01887-5.

Ed summary: A new velocity model improves cell-fate mapping with lineage-traced scRNA-seq data.

PhyloVelo enhances transcriptomic velocity field mapping using monotonically expressed genes

Kun Wang[1,2], Liangzhen Hou[1,3], Xin Wang[1], Xiangwei Zhai[4], Zhaolian Lu[1], Zhike Zi[1], Weiwei Zhai[5,6], Xionglei He[4], Christina Curtis[7,8,9], Da Zhou[2,10*], Zheng Hu[1*]

[1]CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

[2]School of Mathematical Sciences, Xiamen University, Xiamen, China

[3]Faculty of Health Sciences, University of Macau, Taipa, Macau, China

[4]MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China

[5]CAS Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[6]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

[7]Department of Medicine, Division of Oncology, Stanford University School of Medicine, Stanford, CA, USA

[8]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA

[9]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA

[10]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China

* Correspondence to: zhouda@xmu.edu.cn (D.Z.); zheng.hu@siat.ac.cn (Z.H.)

Abstract

Single-cell RNA-sequencing (scRNA-seq) is a powerful approach for studying cellular differentiation, but accurately tracking cell-fate transitions can be challenging, especially in disease conditions. Here, we introduce PhyloVelo, a computational framework that estimates the velocity of transcriptomic dynamics by using monotonically expressed genes (MEGs), or genes with expression patterns that either increase or decrease, but don't cycle, through phylogenetic time. Through integration of scRNA-seq data with lineage information, PhyloVelo identifies MEGs and reconstructs a transcriptomic velocity field. We validate PhyloVelo using simulated data and *C. elegans* ground-truth data, successfully recovering linear, bifurcated, and convergent differentiations. Applying PhyloVelo to seven lineage-traced scRNA-seq datasets, generated via CRISPR/Cas9 editing, lentiviral barcoding or immune repertoire profiling, demonstrates its high accuracy and robustness in inferring complex lineage trajectories, while outperforming RNA velocity. Additionally, we discover that MEGs across tissues and organisms share similar functions in translation and ribosome biogenesis.

Main

Organism development and disease progression both involve serial cell-fate transitions upon repeated cell divisions. Essentially, all cells in an organism are related by a phylogenetic tree where the root represents the zygote, the branches represent cell divisions, and the leaves represent the terminal cells at various phenotypic states (e.g. cell types)[1-4]. To understand how cell fate is determined, it is important to identify the order of

cell-state transitions acting in the lineage tree and the underlying gene regulatory mechanisms that precipitate these transitions[5].

Single-cell RNA sequencing (scRNA-seq) has been a powerful approach to study cellular differentiations[6-10]. However, the transcriptomic trajectories may or may not be equivalent to the true lineage paths of a progenitor population[11-14]. One example is convergent differentiation where distinct progenitors can converge on the same terminal state[15-19]. In this case, similar cellular states do not reflect a closer lineage relationship[13]. Moreover, predicting the fate directions often requires prior knowledge of the initial/terminal cell types[9] or relies on the information of gene expression diversity during development[10], thus limiting their applications to normally differentiating systems[20]. Abnormal development or disease progression often involves noncanonical cell-fate transitions such as dedifferentiation and transdifferentiation[21], while tackling these processes is still challenging with current approaches. RNA velocity[22, 23] provides a powerful framework to predict cellular state transitions by leveraging the internal kinetics of spliced/unspliced RNAs, and can be readily applied to diseased or perturbated conditions. However, the intrinsic high dynamics of RNA kinetics including transcription, splicing and degradation often violates the constant rate assumptions in the model, which can lead to uncertain estimates[23-25]. Taken together, cell-state transitions are challenging to distinguish using transcriptomic data alone[11-14].

The recent use of CRISPR/Cas9 editing to record cell lineages offers an opportunity to reconstruct the cell lineage tree at whole-organism or whole-organ level[26-29]. Importantly, simultaneous analysis of single-cell transcriptomes and lineage tree makes it possible to uncover complex developmental dynamics, as well as the molecular mechanisms, of cell fate commitment[13, 30-35]. For instance, CRISPR lineage tracing has enabled the identification of transcriptional convergence of endodermal cells from both extra-embryonic and embryonic origins during mouse embryogenesis[32]. Although the significance of CRISPR

lineage tracing coupled with single-cell transcriptomics has been widely acknowledged in developmental biology and somatic evolution[13, 29, 33], computationally integrating dual information for reconstructing cellular trajectories is challenging, partially due to the distinct data modalities. Previous effort such as the CoSpar algorithm[36] has been made to use the paired scRNA-seq and lineage information to infer transition maps and predict fate bias of progenitor cells, which is better fit for static barcoding information (e.g. LARRY system[37]). Another algorithm lineageOT has taken advantage of lineage tree for trajectory inference[38], however, this relies on time-course scRNA-seq data and an invariant cell lineage tree. This type of data is currently only available in *C. elegans*[14], thus preventing its application to more common datasets such as CRISPR-based lineage tracing data.

In this study, we described a method to systematically map cell-fate transitions by using both single-cell transcriptomic and lineage information. Our method, called PhyloVelo, leverages monotonically expressed genes (MEGs) along cell divisions to quantify the transcriptomic velocity fields from lineage-resolved single-cell RNA-seq data (**Fig. 1**). We verified the capacity and robustness of PhyloVelo to resolve complex lineage structures in comprehensive simulations and real lineage tracing data in embryo development (*C. elegans* and mouse embryos), tumor evolution (both initiation and metastasis), *in vitro* hematopoiesis and intratumoral T cell dynamics. We further demonstrated that the velocities estimated from one scRNA-seq dataset were sufficiently robust to infer the lineage trajectory with independent datasets in similar biological conditions, even in the absence of lineage information. Finally, we found MEGs were strongly enriched in ribosome-mediated processes across tissues and organisms, thus exposing an internal clock-like gene expression program during cell proliferation and differentiation.

Results

A transcriptomic velocity field reconstructed by MEGs

RNA velocity[22, 23] exploits the kinetics of spliced/unspliced RNAs to estimate the time derivative (or velocity) of single-cell gene expression states ($\mathrm{d}\boldsymbol{s}/\mathrm{d}t$ with $\boldsymbol{s}$ representing the high-dimensional expression state and t representing time). This enables the prediction of future gene expression states and reconstruction of a velocity vector field of cellular state transitions on low-dimensional embedding. We anticipated that other measurements of $\mathrm{d}\boldsymbol{s}/\mathrm{d}t$ can be similarly employed to establish the velocity field. In particular, lineage tracing by endogenous mutations or evolving barcodes (e.g. CRISPR/Cas9 editing) reconstructs a cell phylogenetic tree that records the cell division history from a common progenitor (**Fig. 1a**). Because more cell divisions usually indicate more advanced differentiation stages in a stem cell hierarchy, the phylogenetic time potentially associates with the differentiation time of individual cells. To bridge differentiation and phylogenetic time, we focused on a group of genes (namely MEGs) whose expressions increase or decrease monotonically over phylogenetic time (**Fig. 1a-b**). The monotonic feature of MEGs enables them to serve a "clock" of cell differentiations (**Fig. 1a**).

To identify MEGs and estimate their expression velocity $\mathrm{d}\boldsymbol{s}/\mathrm{d}t$, we first sought to estimate the latent gene expression of each gene at a phylogenetic time, which was akin to the latent variables in single-cell RNA-seq denoising[39]. Inspired by the classic models of trait evolution in phylogenetics[40], we modeled the continuously varying gene expressions by a diffusion process (also called stochastic differential equation) (**Methods**). Each gene has a specific rate of expression dynamics, namely the drift coefficient $v(t, z_t)$ per unit of time, where t is the time started from the root of tree and z_t is the latent expression for a gene at time t. MEGs were identified by their significant association (Pearson's correlation, q<0.05) between the latent expressions, $\boldsymbol{Z} = (\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n)$, and the time from terminal cells to the root, $\boldsymbol{B} = (b_1, b_2, \ldots, b_n)$, where n was the cell number (**Methods**). It's worth noting that MEGs are defined with respect to the observed phylogenetic time range $\boldsymbol{B}$**,** which means the monotonic expression might not retain out of $\boldsymbol{B}$. Moreover, although we focused on

linear MEGs with phylogenetic time, non-linear MEGs can be also identified if the linear regression is statistically significant. We will show later by simulations that linear approximation is technically sound way to accurately map cell trajectories. The drift coefficients of all G MEGs in a dataset, $\boldsymbol{v} = (v_1, v_2, \cdots, v_G)$, were thus referred to as phylogenetic velocity (or PhyloVelo). As shown in **Fig. 1b-c**, phylogenetic velocity can be used to predict the past expression state of each cell before a unit of time Δt (one cell division or mutation), $\boldsymbol{s}^* = \boldsymbol{s} - \boldsymbol{v}\Delta t$. Similar to RNA velocity[22, 23], phylogenetic velocity $\boldsymbol{v}$ can also be projected into low dimensional embedding such as t-distributed Stochastic Neighbor Embedding (tSNE) or Uniform Manifold Approximation and Projection (UMAP), which reconstructs the velocity vector fields (**Fig. 1d**). Unlike RNA velocity where the velocity fields point to future extrapolated states, phylogenetic velocity fields point to the instantaneously past states, thus reconstructing fate-transition map in backward directions (**Fig. 1d**).

PhyloVelo recovers complex lineages in simulations and C. elegans

We next sought to test PhyloVelo with simulation data where various lineage structures were considered, including linear, bifurcated and convergent differentiations (**Fig. 2a-c**). A lineage-imbedded scRNA-seq simulator PROSSTT[41] was modified to record individual cell divisions and generate single-cell UMI counts simultaneously (**Methods**). To model cell differentiations, different cell types each showing a characteristic gene expression program were simulated in the three lineage structures, respectively (**Fig 2a-c, Supplementary Fig. 1**). We also simulated random mutations that occur during cell divisions, which allows to build mutation-based cell lineage tree (**Methods**). Of note, simulations showed that different cell types were highly intermixed on the lineage tree (**Fig. 2a-c**), a phenomenon that appears to be common for organ development across diverse species, such as flies[42], zebrafish[43] and mice[44]. Nevertheless, the dimensionality reduction embedding (tSNE) of simulated scRNA-seq data recapitulated the actual lineage structures (**Fig. 2a-c**).

By applying PhyloVelo to the simulation data, we first found that MEGs following either increasing or decreasing dynamics can be robustly detected with our algorithm (**Extended Data Fig. 1**). With the estimated phylogenetic velocities of MEGs, PhyloVelo mapped the state transitions in backward directions which point to the extrapolated past states of individual cells on low dimensional embedding (**Fig. 2d-f**). We used two quantitative metrics to systematically evaluate the performance of PhyloVelo with simulation datasets generated under a variety of parameters and conditions: (1) precision rate of MEG identification, M_1; (2) Pearson's correlation of estimated velocity directions using the identified MEGs vs using the genuine MEGs, M_2. As shown in **Extended Data Fig. 1** and **Supplementary Fig. 2-5,** we found PhyloVelo inferences were highly robust to the cell number (mean M_1>80% and mean M_2>90% even at low cell number of 100 cells), non-linear dynamics of MEGs (mean M_1>75%, M_2 was not available here), the methods of dimensionality reduction embedding (mean M_2>90% for both tSNE and UMAP), and the sparsity level of single-cell data (mean M_1>90% and M_2>90% even for ~0.2 UMIs per cell per gene). Interestingly, both M_1 and M_2 remained high when the number of genuine MEGs exceeded 50 (M_1>85% and M_2>90%, **Extended Data Fig. 1, Supplementary Fig. 6**). Mathematical analysis verified a small angle (upper bound<37°) between the estimated and true velocity vectors at 50 MEGs and a precision rate of M_1=80% (**Supplementary Note**). Using a stringent threshold (M_2=95%) for good performance, our simulations revealed at least 35, 82 and 43 MEGs were required for linear, bifurcated and convergent differentiation model, respectively. In summary, our comprehensive benchmarking and mathematical analysis demonstrated the high accuracy and robustness of PhyloVelo to systematically map cell-state trajectories.

In addition, we found the estimated phylogenetic velocities based on the mutation-based phylogenies and the ground-truth division history were highly concordant, although inaccurate velocity estimations in local lineages were noted when the mutation rate was rather low (mean mutation rate u=0.1 per cell division) (**Supplementary Fig. 7-8**).

Importantly, phylogenetic velocity estimates were robust to different phylogenetic methods (e.g. maximum likelihood, neighbor joining or maximum parsimony), which was because these methods gave highly consistent inferences on the phylogenetic distances (**Supplementary Fig. 9**). Finally, while classic trajectory inference algorithms such as monocle3[7], slingshot[45], and PAGA[46] can accurately identify the backbones of linear and bifurcated lineage structures, only PAGA was able to identify the circular structure in convergent differentiation (**Supplementary Fig. 10**). In fact, additional information on initial or terminal cell types is needed to define the directions using the aforementioned three algorithms. This is expected because most trajectory inference methods are inadequate for single-cell datasets containing a convergent trajectory, and also rely on prior information of initial/terminal cell types[9].

We next applied PhyloVelo to *C. elegans* given that the embryonic lineage tree of this organism is entirely known[2]. The scRNA-seq data from temporal *C. elegans* embryos are also available and have been mapped to the invariant lineage tree, as described by Packer *et al*.[14]. Thus, *C. elegans* is an ideal system to benchmark our method. We focused on the AB lineage with mostly ectoderm accounting for ~70% of the terminal cells in the embryo (**Fig. 3a**), which also had the densest single-cell annotations in Packer *et al*. dataset[14] spanning from generation 5 (32-cell stage) to 12 (threefold stage of development). Since many nodes on the lineage tree have been sampled multiple times through pooled sequencing of multiple embryos, one cell was randomly chosen to represent the corresponding lineage node. This resulted in 298 non-repetitive cells for the AB lineage, denoted as a single pseudo-embryo. By analyzing the correlation between latent gene expressions and cell generation times, we identified 326 significant (q<0.05) MEGs with 22 and 304 increasing and decreasing in expressions, respectively (**Fig. 3b, Supplementary Table 1**). This was consistent with the observed global decline in gene expressions during *C. elegans* embryogenesis[14].

To generate a ground-truth velocity field, each cell was assigned a vector on the UMAP plot that points to its immediate parental cell in the ground-truth lineage tree (**Fig. 3c**). The vector fields together tracked the cell lineages back to the earliest cells in development. Therefore, by comparing the quantitative directions of PhyloVelo velocity fields with the ground-truth and also RNA velocity fields, we were able to evaluate the accuracy of our method. The UMAP embedding clearly reflected the differentiation trajectories along cell divisions (**Fig. 3c-e**) or embryo time (**Fig. 3f**). Surprisingly, we found that RNA velocity (scVelo - dynamical mode, **Fig. 3d**) failed to recover the expected trajectories, where the directions were even reversed from the ground truth (**Supplementary Fig. 11a**). In fact, the scVelo latent time (**Fig. 3g**) was negatively correlated with the real embryo time in early development before ~300 minutes (**Fig. 3d, Supplementary Fig. 11b**). In contrast, the directions of phylogenetic velocities recapitulated the actual development orders (**Fig. 3e, Supplementary Fig. 11c-d**). Other single pseudo-embryo data also showed similar results (**Supplementary Fig. 12**).

The RNA velocity fields estimated by pooling all 29,600 AB lineage cells from multiple embryos were improved (**Fig. 3i-j**), suggesting that RNA velocity estimates had been hindered by a small cell number. Remarkably, the phylogenetic velocities of 326 MEGs estimated from single pseudo-embryo data (~300 cells) can be used to accurately infer the velocity fields for all 29,600 AB lineage cells, even though their lineage trees were not utilized (**Fig. 3k-l**). In fact, these MEGs were even applicable to non-AB lineage cells such as hypodermis, body wall muscle (BWM) and pharynx (**Extended Data Fig. 2**). Interestingly, a convergent trajectory for the first row of head body wall muscle (BWM) and all other BWMs (including C, D and MS lineages) can be identified (**Extended Data Fig. 2e**). These results demonstrated a general transcriptomic clock during *C. elegans* embryogenesis, and also suggest that compiling a reference panel of MEGs will greatly facilitate the applications of PhyloVelo to conventional scRNA-seq data where lineage data

is unavailable. In summary, the benchmarking on comprehensive simulations and *C. elegans* embryo lineages demonstrated the high robustness of PhyloVelo to recover complex developmental trajectories with phylogeny-resolved scRNA-seq data even with relatively limited cell numbers.

PhyloVelo resolves multiple-rate kinetics in mouse embryos

We next applied PhyloVelo to a CRISPR/Cas9-based lineage tracing dataset from mouse early embryos (E8.0 or E8.5), described by Chan *et al*.[32]. This study provided both cell lineage tree and scRNA-seq data via CRISPR lineage tracing of mouse fertilization through gastrulation. By analyzing four embryos (embryo 1, 2, 3 and 6, each with 6,328 to 19,071 cells and more than 500 unique barcode alleles), we have identified 426, 460, 420 and 418 MEGs ($q<10^{-5}$), respectively at whole embryo level (**Extended Data Fig. 3, Supplementary Table 1**). Notably, about 50% (n=212) of MEGs were overlapped by all four embryos and the phylogenetic velocities of these overlapped MEGs were strongly correlated (Pearson's r=0.65-0.95, **Extended Data Fig. 3**). Given the generally low barcode diversity in CRISPR lineage tracing[47] and also high noise in scRNA-seq data, these data actually indicated the high robustness of PhyloVelo for identifying MEGs. Because of the rapid cell replication in early embryogenesis, the MEGs identified from one snapshot sample from Chan *et al*. dataset[32] might only represent a short-term monotonic effect. Nevertheless, we found about a half (104 out of 212) of overlapped MEGs identified from Chan *et al*. dataset[32] also showed significant correlation ($p<0.05$) with the capture time (E6.5-8.5) of temporal mouse embryos from Pijuan-Sala *et al*.[19] (**Supplementary Fig. 13**). We thus called these 104 genes long-term MEGs, or LT-MEGs. As expected, these 104 LT-MEGs identified from Chan *et al*. dataset[32] enabled accurate prediction of the entire differentiation trajectories of mouse embryogenesis with the temporal scRNA-seq data (**Extended Data Fig. 4a-c**). Remarkably, these LT-MEGs were also highly robust to infer the velocity fields when

transferred to mouse brain tissues across broader developmental stages (E7-18) and over 18 cell types[48] (**Extended Data Fig. 4d-f**).

To directly compare PhyloVelo with RNA velocity, and also quantify the state-transition probabilities between cell types, we next focused on the erythroid lineage given its well-defined differentiation trajectory during mouse gastrulation[49]. Embryo 3 (E8.5) had the largest cell number and more diverse cell types in erythroid developmental lineages (n=2,419 cells), thus being selected for a representative case while other embryos (1, 2 and 6) were also analyzed (**Fig. 4a, Extended Data Fig. 5**). RNA velocity failed to identify hematopoietic/endothelial progenitors as the earliest cell types (**Fig. 4b-c**). In addition, the fractions of varying cell types only changed slightly along the scVelo latent time (**Fig. 4d**). In contrast, PhyloVelo correctly predicted the expected trajectory from hematopoietic/endothelial/primitive blood progenitors to primitive blood early/late based on the velocity fields and pseudotime (**Fig. 4e-g**). Transferring the MEGs identified from erythroid cells of embryo 3 to other three embryos also robustly recovered their erythroid differentiation orders (**Extended Data Fig. 5b-d**). Dynamo[50] was further used to incorporate PhyloVelo velocity fields, which can quantify the transition probabilities between any two cell types (**Extended Data Fig. 5e-h**). Dynamo successfully placed hematopoietic endothelial progenitors and primitive blood late as the starting and ending states, respectively (**Extended Data Fig. 5i-l**). Importantly, the possible ancestral states of a particular cell type (non-zero transition probabilities) recapitulated well where the cell type was differentiated. Finally, by applying CellRank[20] with the input of the PhyloVelo pseudotime, we were also able to identify the known driver genes underlying erythroid maturation (e.g. *Alas2*, *Bpgm*, *Car2*, *Slc4a1*, *Hemgn*, **Supplementary Fig. 14**).

Studies have shown that multiple-rate kinetics (MURK) of RNA violates the constant assumptions in RNA velocity model, which might lead to erroneous estimates of velocities[24, 25]. Erythroid development is a salient example, where due to MURK, the directions of RNA

velocity were even reversed from the expected trajectory[19, 24] (**Fig. 4h**). Remarkably, using the phylogenetic velocities of MEGs in erythroid development from a single embryo (embryo 3) of Chan *et al*. dataset[32] (**Supplementary Fig. 15**), PhyloVelo accurately predicted the expected erythroid trajectory with the scRNA-seq data of temporal mouse embryos (E6.5-8.5) from Pijuan-Sala *et al*.[19], despite the lineage tree was not being available (**Fig. 4i**). The PhyloVelo pseudotime was also strongly correlated with mouse embryo time (**Fig. 4j-k**). Together, these data demonstrated that PhyloVelo can circumvent the MURK issue of RNA velocity and the MEGs identified from one dataset can be also applied to independent datasets, even when phylogenetic information is not available.

PhyloVelo identifies lung tumor dedifferentiation

We next applied PhyloVelo to a CRISPR/Cas9-based lineage tracing dataset in a genetically-engineered mouse model (GEMM) of lung adenocarcinoma (*Kras*$^{LSL\text{-}G12D/+}$; *Trp53*$^{fl/fl}$, or KP model), described by Yang *et al.*[51]. Cancer GEMMs allow one to study tumor evolutionary trajectory in its native microenvironment. Two primary tumors from KP mice (3726_NT_T1 and 3435_NT_T1) were selected because of their relatively high resolution of the lineage trees and composition of diverse cell types (including AT2-like, AT1-like, Gastric-like, High plasticity, Lung-mixed, Endoderm-like, Early EMT (epithelial–mesenchymal transition)-1, etc.) (**Fig. 5a, Extended Data Fig. 6a**). In total, 337 and 344 MEGs (q<0.05) were identified from these two tumors, respectively (**Supplementary Fig. 16, Supplementary Table 1**). RNA velocity by scVelo performed reasonably well in 3435_NT_T1 to recapitulate the expected trajectory from AT2-like to High plasticity, and to Lung-mixed cells (**Extended Data Fig. 6b**), whereas no clear trajectory was inferred in 3726_NT_T1 by scVelo (**Fig. 5b**). In contrast, in both tumors PhyloVelo identified AT2-like cells as the cell-of-origin of KP lung adenocarcinoma and also recovered the trajectory from AT2-like to lung-mixed or Early EMT (**Fig. 5c, Extended Data Fig. 6c**). In 3726_NT_T1, two trajectories appeared to coexist, namely 1) AT2-like > lung-mixed > Early EMT and 2) AT2-

like > Endoderm-like > Early EMT (**Fig. 5c**), thus recapitulating the findings in the original study[51]. As previous reports[52, 53], lung tumor development was accompanied by the loss of AT2 identidy and gain of highly plastistic phenotypes such as lung-mixed and EMT.

Yang *et al.*[51] defined a single-cell fitness signature (**Fig. 5d, Extended Data Fig. 6d**) where the expression of a specific gene module is associated with the cell proliferating fitness estimated from the phylogenetic tree. While no overt association between scVelo latent time and the fitness signatures was found (**Fig. 5d, Extended Data Fig. 6e**), the PhyloVelo pseudotime showed a strong correlation with the fitness signature in both tumors (3726_NT_T1, Spearman's ρ=0.86, P=1.2x10^{-218}; 3435_NT_T1, Spearman's ρ=0.83, P=4.0x10^{-280}, **Fig. 5d**, **Extended Data Fig. 6f**). This indicates an intrinsic link between our measure of phylogenetic velocity and the cell fitness. Interestingly, CytoTRACE[10], a computational algorithm to predict the cellular differentiation states with scRNA-seq data, revealed a drastic increase of the expressed gene number during the tumor evolution (**Fig. 5e-f**, **Extended Data Fig. 6g-i**), which was in line with a dedifferentiation model. Reanalysis of a scRNA-seq dataset from human non-small cell lung cancers[54] verified that CytoTRACE scores increased as tumor evolved from normal lung tissue, early-stage cancer, advanced-stage cancer to pleural fluids and lymph node metastasis (**Fig. 5g**). This suggested that dedifferentiation might be a general phenomenon during tumor evolution. Also, this indicated that although gene expression diversity is a key feature of the developmental potential[10], the directions of cell-state transitions are highly context-dependent and the cellular trajectories in normal differentiation and disease progression can be completely reversed.

Again, we showed that the phylogenetic velocity of MEGs identified from only one pilot tumor 3726_NT_T1 (754 cells, **Supplementary Fig. 16**) enabled the robust inference of velocity fields for other independent KP tumors even the lineage trees were not utilized (n=58,022 cell, **Fig. 5h-i**). In fact, the inferred and expected trajectories were highly

consistent for the cell types that existed in 3726_NT_T1 but not for the cell types such as Mesenchymal 1 and 2 (**Fig. 5h-i**). Interestingly, the phenomenon that two trajectories coexist as in 3726_NT_T1 (**Fig. 5c**) was more evident on the pooled PhyloVelo velocity fields (**Fig. 5h**) and the quantitative state-transition map computed by Dynamo based on the PhyloVelo velocity fields (**Fig. 5j-k**). These results demonstrated the generality of transcriptomic clock across the KP tumors, but also implied that a large single-cell lineage tree spanning numerous cell types must be reconstructed in order to identify more ubiquitously clock-like MEGs.

PhyloVelo for clonal lineage tracing data using static barcodes

Clonal lineage tracing by static barcoding has been paired with single-cell transcriptomics, such as LARRY[37], CellTagging[55] and immune repertoire profiling[56] (e.g. scVDJ-seq), which provides both clonality and gene expression profiles of individual cells. Clonal lineage tracing identifies cells of common ancestry but can't resolve phylogenetic relationship within each clonal subpopulation. Nevertheless, similar to mutation number, the clone sizes (number of cells sharing a unique static barcode) also indicate the relative proliferative activity of the cells in the past division history. Hence, we considered using clone size as a surrogate of phylogenetic time in PhyloVelo (**Fig. 6a**). According to a simple exponential growth model ($c_t = c_0 e^{rt}$), the logarithm of clone size ($\log(c)$) has a linear relationship with cell proliferation rate (r). Therefore, here "MEGs" can be identified by the significant association of latent expressions $\boldsymbol{Z} = (\hat{z}_1, \hat{z}_2, \dots, \hat{z}_n)$ with the logarithm of clone size $\boldsymbol{B} = (\log(c_1), \log(c_2), \dots, \log(c_n))$ (**Fig. 6b**). The velocities were estimated the same way as using phylogeny-resolved scRNA-seq data (e.g. CRISPR lineage tracing) (**Methods**).

We first applied this extended model of PhyloVelo to a lentiviral barcoding dataset from *in vitro* hematopoiesis[37]. This dataset sampled hematopoietic differentiation over the culture of 2, 4, and 6 days and contained 29,242 cells where each could be traced by one unique

362 barcode (**Fig. 6c**). In total, 419 MEGs (q<0.05) were identified, with 297 positively and 122
363 negatively associated with the logarithm of clone sizes at day 6 (**Supplementary Fig. 17a,**
364 **Supplementary Table 1**). PhyloVelo velocity fields accurately traced differentiated cells
365 (erythrocytes, megakaryocytes, mast cells, neutrophils, monocytes, *etc*) backward to
366 undifferentiated progenitor cells (**Fig. 6d-e**). In fact, PhyloVelo pseudotime was strongly
367 correlated with the clonal fate potency inferred by the Cospar algorithm[36] (**Supplementary**
368 **Fig. 18**), indicating clone size-based PhyloVelo has successfully recovered the
369 hematopoietic differentiation trajectories.
370
371 We also showcased the application of PhyloVelo to immune repertoire profiling data, where
372 simultaneous lineage receptor sequences and gene expression profiles of individual T cells
373 are available. With a lineage tracing dataset of intratumoral CD8+ T cells in basal cell
374 carcinoma[57] (**Supplementary Fig. 17b**), PhyloVelo combined with Dynamo[50] quantified the
375 T cell state-transition rates pre and post PD-1 blockade treatment (**Fig. 6f-k)**. The
376 quantitative transition map revealed that the enriched CD8+ activated T cells post treatment
377 had few origin (2.4%) from the infiltrated naïve or memory CD8+ T cells (**Fig. 6j**), in line with
378 the clonal replacement model of T lymphocytes after PD-1 treatment[57]. Interestingly, the
379 hybrid activated/exhausted CD8+ T cells appeared to be mainly (81%) derived from
380 exhausted CD8+ T cells before PD-1 treatment (**Fig. 6h-i**), while they were instead almost
381 all (99%) from activated T cells after treatment (**Fig. 6j-k**). Therefore, our quantitative
382 analyses revealed the drastic fate plasticity of intratumoral CD8+ T cells during checkpoint
383 blockade immunotherapy.

384 ***Comparison of PhyloVelo with different RNA velocity methods***

385 We noticed several methods for estimating RNA velocity have been developed, which
386 model cell-specific and/or gene-specific RNA kinetics with deep learning framework such as
387 VeloVAE[58], DeepVelo[59] and cellDancer[60], or by the radial basis function such as

UniTVelo[61]. These methods highlighted their improved performance as compared to scVelo. We therefore sought to compare PhyloVelo with these RNA velocity estimators using the scRNA-seq datasets of *C. elegans*, mouse erythroid cells and KP mouse lung tumor in this study. For *C. elegans*, only cellDancer seemed to improve the RNA velocity estimates relative to scVelo, where all others still gave reversed directions against the *C. elegans* embryo time (**Supplementary Fig. 19**). For mouse erythroid development (E8.5), UniTVelo showed the best performance amongst the five RNA velocity methods with competitively accurate estimations as PhyloVelo (**Extended Data Fig. 7**). Finally, for KP mouse lung tumor (3726_NT_T1), DeepVelo and UnitVelo performed reasonably well to recapitulate the expected trajectory from AT2-like to Lung-mixed cells. PhyloVelo pseudotime still showed the best correlation with cell fitness signatures (**Supplementary Fig. 20**). Overall, these preliminary comparison analyses highlighted the superior performance of PhyloVelo relative to RNA velocity methods.

MEGs are enriched in ribosome-mediated processes

To systematically investigate the potential functions of MEGs across tissues and organisms, we analyzed three additional CRISPR-based lineage tracing datasets that were derived from mouse or human cell lines, including pancreatic cancer KPCY[62], lung cancer A549[63] and normal epithelial cells HEK293T[64]. The KPCY and A549 cells were sampled from *in vivo* mouse xenograft model, while the HEK293T cells were from a single-cell derived clone of *in vitro* culture. Interestingly, although these cell lines are known to be non-differentiating, continuous cell-state transitions were evident according to the PhyloVelo velocity fields (**Extended Data Fig. 8, Supplementary Figs. 21-22**). For instance, PhyloVelo recovered a dynamic EMT trajectory during the metastatic progression of KPCY and A549 cells in mouse xenografts (**Extended Data Fig. 8, Supplementary Figs. 21**). Even for *in vitro* culture of HEK293T cells, PhyloVelo and scVelo consistently showed continuous state transitions (**Supplementary Fig. 22**), and this phenomenon was not caused by cell-cycle

heterogeneity (**Extended Data Fig. 9**). Interestingly, the CytoTRACE "stemness" scores were strongly associated with PhyloVelo pseudotime in A549 mouse xenografts (**Supplementary Fig. 23**), also in line with a dedifferentiation process during the *in vivo* tumor progression.

We found the MEGs identified across organisms (mouse and human) and tissue or cell types (embryo, tumor tissues, cell lines and intratumoral T cells) were significantly overlapped (**Extended Data Fig. 10a-b**). Interestingly, the ribosome machinery was strongly enriched across the tissues and organisms, including translation, ribonucleoprotein complex biogenesis, ribosome biogenesis and assembly (**Fig. 6l**). For instance, the gene scores of ribosomal protein (RP) genes were significantly associated with the phylogenetic time based on tree distance in KP lung tumors or based on clone size in *in vitro* hematopoiesis and Intratumoral CD8+ T cells (**Supplementary Fig. 24**). To rule out the possibility that ribosomal genes were identified because of their high expression heterogeneity amongst the cells, we further performed permutation analysis where the phylogenetic distances were randomly shuffled and assigned to the cells. Here, although some "pseudo-MEGs" can still be identified (**Extended Data Fig. 10c**), they only showed weak associations (most *q* values were around 0.05) with the phylogenetic distances. Importantly, no significant enrichment in ribosome-mediated processes was found (**Extended Data Fig. 10d**). These results strongly suggest that many ribosomal genes genuinely follow the clock-like expression dynamics during cell proliferation and differentiation.

Discussion

Defining the correct directions of cell-fate transitions is crucial for unraveling the (epi)genetic regulators that drivers of lineage specification in diverse biological contexts[20]. Although RNA velocity and its improvements[22, 23, 50, 61] are powerful approaches for quantifying cellular

transitions from single-cell transcriptomic data, an accurate estimation of the velocity fields is still challenging because of the highly dynamic RNA kinetics (transcription, splicing and degradation)[23-25] and the biased capture of intron regions by droplet-based scRNA-seq[65]. The fundamental objective of our PhyloVelo algorithm is the same as RNA velocity – to extrapolate the gene expression of single cells to their near future or past states. However, unlike RNA velocity, PhyloVelo quantifies the transcriptomic velocity by measuring the rate of expression changes along a cell division history. Using various single-cell datasets where the coupled lineage information was available, PhyloVelo not only recovered the expected trajectories more accurately, but also gave more consistent estimates of the velocities relative to RNA velocity, across diverse biological contexts.

Analysis of lineage-resolved scRNA-seq datasets with PhyloVelo across mouse embryo development, hematopoietic differentiation, tumor evolution and immune cell dynamics yields insights into cell state dynamics. First, in each of the lineage tracing datasets, we have identified 100-500 MEGs, suggesting a considerable number of genes follow directional expression trajectories along cell divisions, at least within the phylogenetic time range of sampled cells. Interestingly, the MEGs across tissues and organisms had highly similar functions in translation and ribosome biogenesis, in line with their crucial role in regulating cell proliferation. Previous studies have also shown that ribosomal protein (RP) genes are commonly downregulated during differentiation[66, 67], which represent robust markers of differentiation potency[68, 69]. Our study provides an explanation on why they serve as markers of differentiation potency - that is probably through regulating cell cycle and proliferation. In other words, the downregulation of some RP genes might suppress cell proliferation and thus promote differentiation. Second, we showcased that the phylogenetic velocities of MEGs estimated from one lineage-resolved scRNA-seq dataset can be reused in independent scRNA-seq datasets in similar biological conditions, in the absence of lineage information. Because obtaining a coupled lineage tree for every scRNA-seq dataset

is rather laborious, the transferability of MEGs facilitates the application of PhyloVelo to conventional scRNA-seq datasets. It is important to note that the transfer of MEGs is limited to similar biological conditions. However, it is not recommended to transfer MEGs between different conditions, such as normal and disease as the MEGs can differ significantly. To study normal development, it would be beneficial for future efforts to compile a comprehensive set of MEGs encompassing whole-organism development by employing whole-organism lineage tracing. These MEGs can then be used to assess their phylogenetic velocities in various organs or tissues. On the other hand, in the context of diseased conditions, it is essential to identify specific MEGs for each dataset, utilizing the corresponding lineage tracing data. Third, by combining PhyloVelo and Dynamo[50], we were able to estimate the transition probability between any two cell states. For instance, in mouse lung tumors, we found two competing trajectories of cell-state evolution through dedifferentiation. In another case of intratumor CD8+ T cells, we found distinct origin of activated T cells pre and post anti-PD-1 treatment. These quantitative analyses revealed high cell plasticity for both tumor cells and immune microenvironment during tumor progression and treatment. Importantly, as RNA velocity, PhyloVelo velocities fields are useful for identifying cell-fate drivers[20] or core gene regulatory networks[70].

Despite the rapid development of CRISPR lineage tracing methods[4, 29, 71], building high-precision lineage trees with single-cell resolution is still challenging because of the small number of Cas9 target sites (typically<50), rapid saturation, frequent inter-site deletions, and other factors[47, 72, 73]. This requires a more reliable lineage tracing method that has a larger lineage-labeling space and more stable mutagenesis strategy. We recently developed a base editor-based lineage tracing method, called SMALT[42], which leverages a genetically-evolved activation-induced cytidine deaminase (AID) to specifically target a 3k synthetic DNA barcode and induce C to T mutations on it with high efficacy. The lineage tree reconstructed by SMALT achieved nearly single-cell resolution and over 80% statistically

bootstrapping support[42]. We envision the combination of SMALT lineage tracing and single-cell transcriptomics will greatly empower PhyloVelo to resolve complex lineage dynamics in more diverse biological contexts, such as genetic perturbation or disease progression.

In summary, we provide a theoretical framework to quantify cell-fate transitions by leveraging both single-cell lineage and transcriptomic information. With the rapid development of single-cell lineage tracing technologies and emergence of lineage-traced multi-omic data[74], we envision our method will facilitate the lineage analysis for complex cellular processes and the discovery of the cell-fate determinants in diverse organisms, tissues, and diseases.

Acknowledgments

We thank Yuanhua Huang, Jiguang Wang, Jin Xu, Liang Ma, Wanze Chen and members of Hu laboratory for constructive discussions. This work was supported by National Key R&D Program of China (2021YFA1302500 to Z.H.), National Natural Science Foundation of China (11971405 to D.Z., 32270693 to Z.H.), Guangdong Basic and Applied Basic Research Foundation (2021B1515020042 to Z.H.), Fundamental Research Funds for the Central Universities (20720230023 to D.Z.) and China Postdoctoral Science Foundation (2021M693303 to Z.L, 2022M723301 to X.W.).

Author contributions

Z.H. and K.W. conceived the concept of phylogenetic velocity. Z.H., K.W., and D.Z. designed the study. K.W. developed the mathematical framework and implemented the software. K.W., Z.H., L.H., Z.L., X.W., X.Z., Z.Y. analyzed the data. W.Z. and Z.Z. provided constructive suggestions on the model. K.W., Z.H., D.Z., C.C., X.H., interpreted results. Z.H. and K.W. wrote the manuscript with contributions from all co-authors. Z.H., and D.Z. supervised the study.

Competing interests

C.C. is an advisor and stockholder in Grail, Ravel, and DeepCell and an advisor to Genentech, Bristol Myers Squibb, 3T Biosciences, and NanoString. Other authors declare no competing interests.

Figure Legends

Fig. 1. Schematic of the PhyloVelo framework. (**a**) Schematic of monotonically expressed genes (MEGs) over phylogenetic time on a cell phylogenetic tree. (**b**) Two examples of MEGs whose latent expressions are associated with the phylogenetic time (cell divisions or mutation number). A diffusion process of gene expressions was used to model the changes of latent expressions over phylogenetic time. This enables the estimation of the phylogenetic velocity, $v = (v_1, v_2, \cdots, v_G)$, which corresponds to the drift coefficients of G MEGs in the diffusion process (approximate to the slope of linear regression between latent expression and phylogenetic time). Whiskers: minimum and maximum. (**c**) Phylogenetic velocity predicts the past transcriptional state of a cell before a unit of phylogenetic time (one cell division or mutation). (**d**) Projection of the phylogenetic velocity into low dimensional embedding enables the mapping of cell-state trajectory in backward directions.

Fig. 2. PhyloVelo recovers complex cell lineages in simulations. Simulation of single-cell RNA-seq data and paired cell-division history under linear (**a**), bifurcated (**b**), and convergent (**c**) differentiation models, respectively. Colors are labeled by cell types. Each simulation consists of 1,000 cells randomly sampled from a growing cell population at 10,000 cells. Each cell has 2,000 expressed genes, including 200-300 MEGs. (**d-f**) Phylogenetic velocity fields reconstructed by PhyloVelo for the corresponding differentiation scenarios. The left panel shows the single-cell level of velocity fields, while the right panel

shows the same velocity fields visualized as streamlines in scVelo. PhyloVelo velocity fields are at backward directions.

Fig. 3. PhyloVelo reconstructs the embryonic differentiation trajectories of *C. elegans*. (**a**) Phylogenetic tree of the *C. elegans* AB lineage. (**b**) Heatmap showing the expressions (z-score normalized) of MEGs along *C. elegans* embryo time. (**c**) The ground-truth velocity fields represent vectors superimposed on the cells that point to their immediate parental cells on the Uniform Manifold Approximation and Projection (UMAP) plot. (**d-e**) The velocity fields estimated by scVelo (dynamical mode) (**d**) or PhyloVelo (**e**). Dash square indicates the early embryonic lineages where RNA velocity gave erroneous estimations on the fate directions. (**f**) *C. elegans* embryo time as Packer *et al.*[14]. (**g**) scVelo latent time. (**h**) PhyloVelo pseudotime. (**i**) RNA velocity fields for all 29,600 AB lineage cells. Colors are labeled by scVelo latent time. (**j**) The correlation between scVelo latent time and embryo time for all AB lineage cells. (**k**) PhyloVelo velocity fields for all 29,600 AB lineage cells, estimated by the phylogenetic velocity of MEGs in a single embryo (n=298 cells). Cell colors are labelled by PhyloVelo pseudotime. (**l**) The correlation between PhyloVelo pseudotime and embryo time for all AB lineage cells. The Spearman correlation coefficients and *P* values are shown.

Fig. 4. PhyloVelo reconstructs the cellular trajectory of mouse erythroid maturation. (**a**) Phylogenetic tree of the 2,419 erythroid lineage cells (embryo 3, E8.5) in Chan *et al.* dataset[32]. (**b-c**) RNA velocity fields (scVelo - dynamical mode) and the latent time of mouse erythroid development. (**d**) Muller plot showing the fractions of four cell types that change over scVelo latent time. (**e-f**) PhyloVelo velocity fields and the pseudotime of mouse erythroid development. (**g**) Muller plot showing the fractions of four cell types that change over PhyloVelo pseudotime. (**h**) Erroneous estimations of RNA velocity fields on erythroid maturation because of multiple rate kinetics (MURK). Data were from Pijuan-Sala *et al.*[19]. (**i**)

PhyloVelo velocity fields of erythroid maturation for Pijuan-Sala *et al*. dataset while using the MEGs identified from Chan *et al*. dataset. (**j**) PhyloVelo pseudotime of erythroid maturation in Pijuan-Sala *et al*. dataset. (**k**) The correlation between PhyloVelo pseudotime and mouse embryo time (n=12,324 cells). The Spearman correlation coefficient and *P* value are shown here. Whiskers: minimum and maximum; center lines: median.

Fig. 5. PhyloVelo identifies a dedifferentiation trajectory in lung tumor evolution. (**a**) Phylogenetic tree of 754 cells from a KP-mouse primary lung tumor, 3726_NT_T1, in Yang *et al*. dataset[51]. The scRNA-seq data, cell type annotations, and lineage trees were obtained from the original study. (**b**) RNA velocity fields (scVelo - dynamical mode). (**c**) PhyloVelo velocity fields. (**d**) Fitness signatures of individual cells, as defined by Yang *et al*. (**e**) CytoTRACE score of individual cells. (**f**) The correlation between PhyloVelo pseudotime and CytoTRACE scores. The Spearman correlation coefficient and *P* value are shown here. (**g**) CytoTRACE score of single tumor cells from human lung primary sites (tLung and tL/B), pleural fluids (PE), lymph node metastases (mLN), and brain metastases (mBrain), as well as normal tissues from lungs (nLung), as described in Kim *et al*.[54]. Bar, median; box, 25th to 75th percentile (IQR); vertical line, data within 1.5 times the IQR. (**h**) PhyloVelo velocity fields for all 58,022 single cells from pooled KP primary lung tumors, estimated by the MEGs identified from 3726_NT_T1. (**i**) PhyloVelo velocity fields for the cell types that existed in 3726_NT_T1. (**j**) Cell-type transition graph (backward) based on the transition rate matrix between any two cell types (**k**), estimated by Dynamo using PhyloVelo velocity fields as input. The arrows point from the current states to the past states.

Fig. 6. PhyloVelo inference with clonal lineage tracing data and MEGs are enriched in ribosome-mediated processes. (**a**) Schematic of clonal lineage tracing data where static barcodes identify cells of common ancestry. Clone size, denoted by c_k for k clones, represents the number of cells carrying the same unique barcode. (**b**) Two examples of

596 clonal size-based MEGs whose latent expressions are positively or negatively associated
597 with the logarithm of clone sizes, respectively. Whiskers: minimum and maximum. (**c**)
598 scRNA-seq data of in vitro hematopoietic differentiation from Weinreb *et al*.[37], where each
599 cell over the course of 2, 4, and 6 days culture could be traced by one unique barcode. (**d**)
600 The velocity fields estimated by PhyloVelo. (**e**) Cell type transition graph (backward) of in
601 vitro hematopoietic differentiation. (**f**) UMAP of tumor-infiltrating CD8+ T cells in BCC
602 samples pre- and post-PD-1 blockade, colored by anti-PD-1 treatment status. Data were
603 from Yost *et al*.[57] (**g**) The velocity fields estimated by PhyloVelo. (**h-i**) Cell-type transition
604 graph and transition matrix (backward) at pre-treatment. (**j-k**) Cell-type transition graph and
605 transition matrix (backward) at post-treatment. CD8_act: CD8+ activated T cells; CD8_ex:
606 CD8+ exhausted T cells; CD8_ex_act: CD8+ exhausted/activated T cells; CD8_eff: CD8+
607 effector T cells; CD8_mem: CD8+ memory T cells. (**l**) Gene ontology (GO) enrichment of
608 MEGs identified across tissues and organisms. The top and most commonly shared 20
609 biological processes are shown. Ribosome-mediated processed are highlighted.
610
611 **Extended Data Fig. 1. Quantitative metrics for evaluating PhyloVelo's performance on**
612 **simulation data.** Two quantitative metrics with varied cell numbers (**a**), non-linear MEGs
613 (**b**), different dimensionality reduction methods (**c**), varied data sparsity (**d**) and varied
614 numbers of MEGs (**e**). All benchmarks are simulated 50 times independently. Bar, median;
615 box, 25th to 75th percentile (IQR); vertical line, data within 1.5 times the IQR.
616
617 **Extended Data Fig. 2. PhyloVelo velocity fields in three additional lineages of *C.***
618 ***elegans*.** (**a-c**) Hypodermis, body wall muscle (BWM) and Pharynx lineage cells,
619 respectively. Colors are labeled by the estimated embryo time (minutes). (**d-f**) PhyloVelo
620 velocity fields of the three lineages respectively each consisting of 2,000 randomly sampled
621 cells from multiple embryos, which were reconstructed using the MEGs identified from 298
622 AB lineage cells. Colors are labeled by the PhyloVelo pseudotime. (**g-i**) The correlation

between PhyloVelo pseudotime and embryo time for the cells in the three lineages. The Spearman correlation coefficients and *P* values are shown.

Extended Data Fig. 3. High concordance of MEGs identified from 4 mouse embryos (E8.0/8.5) in Chan *et al.*[32] (**a**) Venn diagram showing the overlap of MEGs identified from four mouse embryos in the dataset of Chan *et al*. *P* value, one-sided SuperExactTest multi-set intersection test. (**b-g**) The correlation of phylogenetic velocities $\boldsymbol{v}$ for the overlapped MEGs between any two embryos. The Pearson correlation coefficients and *P* values are shown.

Extended Data Fig. 4. The global differentiation trajectories of whole mouse embryos and brain tissues predicted by LT-MEGs. (**a**) PhyloVelo velocity fields of mouse embryos (E6.5-8.5) mapped by 104 LT-MEGs with the temporal scRNA-seq dataset from Pijuan-Sala *et al*.[19] (**b-c**) UMAP plot colored by PhyloVelo pseudotime (**b**) or sample capture time (**c**). (**d**) PhyloVelo velocity fields of mouse brain (E7-18) mapped by LT-MEGs with the temporal scRNA-seq dataset from La Manno *et al*.[48] (**e-f**) tSNE plot colored by PhyloVelo pseudotime (**e**) or sample capture time (**f**). UMAP or tSNE coordinates were as the original studies.

Extended Data Fig. 5. PhyloVelo velocity fields and quantitative state transitions of mouse erythroid development for four embryos from Chan *et al.*[32] (**a-d**) PhyloVelo velocity fields. (**e-h**) The transition rate (backward) between any two cell types. (**i-l**) cell-type transition graph (backward) visualized based on the cell-type transition rates. PhyloVelo velocity fields were used as the input of Dynamo.

Extended Data Fig. 6. PhyloVelo reconstructs the cellular trajectory of lung cancer evolution in 3435_NT_T1. (**a**) Single-cell phylogenetic tree of primary lung tumor 3435_NT_T1 (n=1,109 cells) from KP ($Kras^{LSL\text{-}G12D/+}$;$Trp53^{fl/fl}$) mouse model. The single-cell

RNA data, cell type annotations and lineage tree were obtained from Yang *et al*.[51] (**b**) RNA velocity fields (scVelo - dynamical mode). (**c**) PhyloVelo velocity fields. (**d**) The fitness signatures of single cells as defined by Yang *et al*. (**e**) The correlation between scVelo latent time and fitness signatures. (**f**) The correlation between PhyloVelo pseudotime and fitness signatures. (**g**) CytoTRACE score of individual cells. (**h**) The correlation between scVelo latent time and CytoTRACE scores. (**i**) The correlation between PhyloVelo pseudotime and CytoTRACE scores. (**j**) The correlation of phylogenetic velocities for the overlapped MEGs between KP primary tumor 3435_NT_T1 and 3726_NT_T1. The Pearson correlation coefficients and *P* values are shown here.

Extended Data Fig. 7. Comparison of PhyloVelo with scVelo, VeloVAE, DeepVelo, CellDancer and UniTVelo respectively on mouse erythroid data. scVelo - RNA velocity fields (**a**), latent time (**b**) and the fractions of different cell types along latent time (**c**). VeloVAE - RNA velocity fields (**d**), latent time (**e**) and the fractions of different cell types along latent time (**f**). DeepVelo - RNA velocity fields (**g**), latent time (**h**) and the fractions of different cell types along latent time (**i**). cellDancer - RNA velocity fields (**j**), pseudotime (**k**) and the fractions of different cell types along pseudotime (**l**). UniTVelo - RNA velocity fields (**m**), latent time (**n**) and the fractions of different cell types along latent time (**o**). PhyloVelo - velocity fields (**p**), pseudotime (**q**) and the fractions of different cell types along pseudotime (**r**). PhyloVelo velocity fields are in backward directions.

Extended Data Fig. 8. The dynamic EMT trajectory in metastatic progression of pancreatic cancer KPCY cells. (**a**) Phylogenetic tree of 601 non-repetitive terminal cells in tumor subclone M1.1 from Simeonov *et al*.[62] Cell colors are labeled by EMT pseudotime as defined in the original study. (**b**) The total UMI count (normalized) of MEGs changing with the phylogenetic distance from the root. (**c**) Heatmap of MEG expressions (z-score normalized) with EMT pseudotime. (**d**) RNA velocity fields (scVelo - dynamical mode). Cell

colors are labeled by EMT pseudotime. (**e**) scVelo latent time. (**f**) The correlation between scVelo latent time and EMT pseudotime. (**g**) PhyloVelo velocity fields. Cell colors are labeled by EMT pseudotime. (**h**) PhyloVelo pseudotime. (**i**) The correlation between PhyloVelo pseudotime and EMT pseudotime. The Spearman correlation coefficients and *P* values are shown here.

Extended Data Fig. 9. Continuous state transitions inferred by PhyloVelo after regressing out cell-cycle effect. (**a-d**) PhyloVelo velocity fields after regressing out cell-cycle dynamics in KPCY, A549 lg1, A549 lg2 and HEK293T, respectively. (**e-h**) The correlation of PhyloVelo pseudotime between original analysis and post regressing out of cell-cycle effect in KPCY, A549 lg1, A549 lg2 and HEK293T, respectively. The Pearson correlation coefficients and *P* values are shown here.

Extended Data Fig. 10. Overlap of MEGs across organisms and tissue/cell types and the permutation analysis of MEG identification. (**a**) The overlap of MEGs identified in different datasets as stratified by mouse vs human. (**b**) The overlap of MEGs identified in different datasets as stratified by normal vs tumor cells. *P* values are by one-sided hypergeometric test. (**c**) The *q* values of MEGs in standard and permutation analysis. Permutation analysis was done by randomly shuffling the phylogenetic distances of the cells, followed by the PhyloVelo inference procedure. The number of detected MEGs in standard and permutation analysis respectively are: n=1,724 and n=941 genes in Embryo E8/E8.5; n=681 and n=445 genes in KP lung tumor; n=424 and n=141 genes in KPCY; n=629 and n=50 genes in A549; n=243 and n=90 genes in HEK293T; n=419 and n=112 genes in *in vitro* hematopoiesis; n=368 and n=270 genes in CD8+T cells. Bar, median; box, 25th to 75th percentile (IQR); vertical line, data within 1.5 times the IQR. (**d**) The GO enrichment of pseudo-MEGs across the seven lineage tracing datasets.

References

1. Salipante, S.J. & Horwitz, M.S. Phylogenetic fate mapping. *Proc Natl Acad Sci U S A* **103**, 5448-5453 (2006).

2. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64-119 (1983).

3. Stadler, T., Pybus, O.G. & Stumpf, M.P. Phylodynamics for cell biologists. *Science* **371**, eaah6266 (2021).

4. Baron, C.S. & van Oudenaarden, A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat Rev Mol Cell Biol* **20**, 753-765 (2019).

5. Moris, N., Pina, C. & Arias, A.M. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet* **17**, 693-703 (2016).

6. Bendall, S.C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014).

7. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).

8. Haghverdi, L., Buttner, M., Wolf, F.A., Buettner, F. & Theis, F.J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845-848 (2016).

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

10. Gulati, G.S. et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science* **367**, 405-411 (2020).

11. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M. & Klein, A.M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci U S A* **115**, E2467-E2476 (2018).

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

13. Wagner, D.E. & Klein, A.M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat Rev Genet* **21**, 410-427 (2020).

14. Packer, J.S. et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365** (2019).

15. Mulas, C., Chaigne, A., Smith, A. & Chalut, K.J. Cell state transitions: definitions and challenges. *Development* **148** (2021).

16. Gerber, T. et al. Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. *Science* **362** (2018).

17. Liu, X. et al. Single-Cell RNA-Seq of the Developing Cardiac Outflow Tract Reveals Convergent Development of the Vascular Smooth Muscle Cells. *Cell Rep* **28**, 1346-1361 e1344 (2019).

18. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361-367 (2019).

19. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).

20. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat Methods* **19**, 159-170 (2022).

21. Gupta, P.B., Pastushenko, I., Skibinski, A., Blanpain, C. & Kuperwasser, C. Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance. *Cell Stem Cell* **24**, 65-78 (2019).

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

23. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408-1414 (2020).

24. Barile, M. et al. Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biol* **22**, 197 (2021).

25. Bergen, V., Soldatov, R.A., Kharchenko, P.V. & Theis, F.J. RNA velocity-current challenges and future perspectives. *Mol Syst Biol* **17**, e10282 (2021).

26. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).

27. Frieda, K.L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111 (2017).

28. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361** (2018).

29. VanHorn, S. & Morris, S.A. Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. *Dev Cell* **56**, 7-21 (2021).

30. Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112 (2018).

31. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* **36**, 442-450 (2018).

32. Chan, M.M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82 (2019).

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

34. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat Biotechnol* **36**, 469-473 (2018).

35. Bowling, S. et al. An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, 1693-1694 (2020).

36. Wang, S.W., Herriges, M.J., Hurley, K., Kotton, D.N. & Klein, A.M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat Biotechnol* **40**, 1066-1074 (2022).

37. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D. & Klein, A.M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367** (2020).

38. Forrow, A. & Schiebinger, G. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nat Commun* **12**, 4940 (2021).

39. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).

40. Butler, M.A. & King, A.A. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am Nat* **164**, 683-695 (2004).

41. Papadopoulos, N., Gonzalo, P.R. & Soding, J. PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes. *Bioinformatics* **35**, 3517-3519 (2019).

42. Liu, K. et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat Methods* **18**, 1506-1514 (2021).

43. Wagner, D.E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987 (2018).

44. Salipante, S.J., Kas, A., McMonagle, E. & Horwitz, M.S. Phylogenetic analysis of developmental and postnatal mouse cell lineages. *Evol Dev* **12**, 84-94 (2010).

45. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

46. Wolf, F.A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**, 59 (2019).

47. Salvador-Martinez, I., Grillo, M., Averof, M. & Telford, M.J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *Elife* **8** (2019).

48. La Manno, G. et al. Molecular architecture of the developing mouse brain. *Nature* **596**, 92-96 (2021).

49. Baron, M.H., Isern, J. & Fraser, S.T. The embryonic origins of erythropoiesis in mammals. *Blood* **119**, 4828-4837 (2012).

50. Qiu, X. et al. Mapping transcriptomic vector fields of single cells. *Cell* **185**, 690-711 e645 (2022).

51. Yang, D. et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905-1923 e1925 (2022).

52. Marjanovic, N.D. et al. Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246 e213 (2020).

53. LaFave, L.M. et al. Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* **38**, 212-228 e213 (2020).

54. Kim, N. et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat Commun* **11**, 2285 (2020).

55. Biddy, B.A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219-224 (2018).

56. Penter, L., Gohil, S.H. & Wu, C.J. Natural barcodes for longitudinal single cell tracking of leukemic and immune cell dynamics. *Frontiers in Immunology* **12**, 788891 (2022).

57. Yost, K.E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat Med* **25**, 1251-1259 (2019).

58. Gu, Y., Blaauw, D. & Welch, J.D. Bayesian Inference of RNA Velocity from Multi-Lineage Single-Cell Data. *bioRxiv*, 2022.2007.2008.499381 (2022).

59. Cui, H., Maan, H., Taylor, M.D. & Wang, B. DeepVelo: Deep Learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. *bioRxiv*, 2022.2004.2003.486877 (2022).

60. Li, S. et al. A relay velocity model infers cell-dependent RNA velocity. *Nat Biotechnol* (2023).

61. Gao, M., Qiao, C. & Huang, Y. UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nat Commun* **13**, 6586 (2022).

62. Simeonov, K.P. et al. Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell* **39**, 1150-1162 e1159 (2021).

63. Quinn, J.J. et al. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371** (2021).

64. Choi, J. et al. A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* (2022).

65. Qiu, Q. et al. Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat Methods* **17**, 991-1001 (2020).

66. Athanasiadis, E.I. et al. Single-cell RNA-sequencing uncovers transcriptional states and fate decisions in haematopoiesis. *Nat Commun* **8**, 2045 (2017).

67. Fei, L. et al. Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development. *Nat Genet* **54**, 1051-1061 (2022).

68. Shi, J., Teschendorff, A.E., Chen, W., Chen, L. & Li, T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Brief Bioinform* (2018).

69. Teschendorff, A.E. & Feinberg, A.P. Statistical mechanics meets single-cell biology. *Nat Rev Genet* **22**, 459-476 (2021).

70. Singh, R., Wu, A.P., Mudide, A. & Berger, B. Unraveling causal gene regulation from the RNA velocity graph using Velorama. *bioRxiv*, 2022.2010.2018.512766 (2023).

71. Hughes, N.W. et al. Machine-learning-optimized Cas12a barcoding enables the recovery of single-cell lineages and transcriptional profiles. *Mol Cell* (2022).

72. Gong, W. et al. Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of C. elegans and M. musculus developmental trees. *Cell Syst* **12**, 810-826 e814 (2021).

73. Espinosa-Medina, I., Garcia-Marques, J., Cepko, C. & Lee, T. High-throughput dense reconstruction of cell lineages. *Open Biol* **9**, 190229 (2019).

74. Jindal, K. et al. Multiomic single-cell lineage tracing to dissect fate-specific gene regulatory programs. *bioRxiv*, 2022.2010.2023.512790 (2022).

75. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* **81**, 2340-2361 (1977).

76. Minh, B.Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).

77. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

78. Chen, W. et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol* **19**, 70 (2018).

79. Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM Journal on Applied Mathematics* **80**, 1336-1355 (2020).

80. Prim, R.C. Shortest connection networks and some generalizations. *The Bell System Technical Journal* **36**, 1389-1401 (1957).

81. Cock, P.J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

82. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021).

83. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

84. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94-98 (2017).

85. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

86. Chan, M.M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82 (2019).

Methods

The mathematical framework of PhyloVelo

The dynamics of the latent expression z for each gene on a phylogeny $\mathcal{T}$ was assumed to follow a diffusion process (also known as the stochastic differential equation, SDE), which varies along cell divisions:

$$\mathrm{d}z_t = v(t, z_t)\mathrm{d}t + \sigma(t, z_t)\mathrm{d}W_t \tag{1}$$

Here, W_t is a standard Brownian motion. In our model, we hypothesized that there is a group of genes G_m whose drift coefficient $v(t, z_t)$ and diffusion coefficient $\sigma(t, z_t)$ are independent of both t and z_t, thus $v(t, z_t) = v$ and $\sigma(t, z_t) = \sigma$. We called them monotonically expressed genes (MEGs). For this type of genes, the dynamics of its latent expression z is thus formulated as:

$$\mathrm{d}z_t = v\mathrm{d}t + \sigma\mathrm{d}W_t \tag{2}$$

and its expectation is given by:

$$\mathbb{E}(z_t) = \mathbb{E}\big(z_{t_0}\big)v(t - t_0) \tag{3}$$

For the observed scRNA-seq measurement x (read or UMI count), we assumed that it is sampled from the negative binomial (NB) distribution or the zero-inflated negative binomial (ZINB) distribution:

$$\mathbb{P}(x|z', \alpha, \psi) = \begin{cases} (1-\psi) + \psi\left(\dfrac{\alpha}{\alpha + z'}\right)^{\alpha}, & x = 0 \\ \psi\dfrac{\Gamma(x+\alpha)}{x!\,\Gamma(\alpha)}\left(\dfrac{\alpha}{z'+\alpha}\right)^{\alpha}\left(\dfrac{z'}{z'+\alpha}\right)^{x}, & x \geq 1 \end{cases} \tag{4a}$$

where z' is the exponential function of latent expression z, α is the scale parameter, and ψ is the zero-inflation parameter. The expectation of the distribution is $z'\psi$. For the negative binomial distribution, $\psi = 1$. We used the likelihood ratio test to verify zero inflation for each gene (**Supplementary Note**).

For the scRNA-seq data after normalization (e.g. using *scanpy.pp.normalize_per_cell* and *scanpy.pp.log1p*), we also provided a Gaussian model of latent expression for normalized data:

$$\mathbb{P}(x|z,\alpha,\psi) = \begin{cases} (1-\psi), & x = 0 \\ \psi \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{(x-z)^2}{2\alpha}\right), & \text{otherwise} \end{cases} \quad (4b)$$

To estimate the latent expression z, we used the maximum a posteriori probability (MAP) estimate. For the ZINB model using raw UMI count data, we estimated z' and then took logarithm to get the estimated latent expression z:

$$\hat{z}_{\text{MAP}}(x) = \log\left(\underset{z'}{\text{argmax}}\big(\mathbb{P}(x|,\alpha,\psi)\mathbb{P}(z')\big)\right) \quad (5a)$$

For the Gaussian model using normalized UMI count data, we directly performed the MAP estimate of the latent expression z:

$$\hat{z}_{\text{MAP}}(x) = \underset{z}{\text{argmax}}\big(\mathbb{P}(x|z,\alpha,\psi)\mathbb{P}(z)\big). \quad (5b)$$

For a MEG *g*, its drift coefficient can be estimated as:

$$v_g = \frac{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{B} - n\overline{\boldsymbol{Z}}\,\overline{\boldsymbol{B}}}{\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} - n\overline{\boldsymbol{B}}^2} \quad (6)$$

Here $\boldsymbol{z} = (\hat{z}_1, \hat{z}_2, \ldots \hat{z}_n)$ represents the estimated latent expressions and $\boldsymbol{B} = (b_1, b_2, \ldots, b_n)$ the phylogenetic distances from terminal cells to the root (n is the cell number in a dataset). The drift coefficients of all G MEGs in a dataset $\boldsymbol{v} = (v_1, v_2, \cdots, v_G)$ were thus referred to be as phylogenetic velocity.

For clonal lineage-resolved scRNA-seq data by lentiviral barcoding or immune cell receptor sequences, $\boldsymbol{B}$ represent the logarithm of clonal sizes of individual cells at the time of sampling, namely $\boldsymbol{B} = (\log(c_1), \log(c_2), \ldots, \log(c_n))$, where c is the size of corresponding clone that a cell belongs to and n the cell number. The estimation of $\boldsymbol{v}$ is the same as using phylogeny-resolved scRNA-seq data.

Simulation of phylogeny-resolved scRNA-seq data

To generate simultaneous single-cell phylogenetic and transcriptomic data *in silico*, a lineage-embedded scRNA-seq data simulator, PROSSTT[41], was modified to account for dividing cell populations, so that the whole cell division history initiated from a single cell can be recorded. The simulation consisted of three parts:

1) Simulate a cell division and differentiation process using the Gillespie algorithm to obtain the cell division history;
2) Given a cell differentiation model (linear, bifurcated or convergent), use the diffusion process to generate gene expression programs;
3) Assign the gene expression programs onto the cell division history in order to obtain the read/UMI count data for each gene in each cell.

***Simulating cell division history and mutation-based phylogeny*.** We used a continuous-time Markov process to simulate cell division and differentiation. In particular, each cell type i has a specific division rate $p_i(t)$ and differentiation rate $q_{ij}(t)$, given as follows:

$$\begin{aligned} p_i(t) &= r_i\left(1 - \frac{1}{1 + \mathrm{e}^{-k_i(t-t_{0i})}}\right) \\ q_{ij}(t) &= p_{ij} r_i \left(\frac{1}{1 + \mathrm{e}^{-k_i(t-t_{0i})}}\right) \end{aligned} \tag{7}$$

where r_i, k_i and t_{0i} are the cell-type specific parameters and p_{ij} is the probability of cell type i differentiating into cell type j, $\sum_{i \neq j} p_{ij} = 1, p_{ii} = 0$.

Now, we can simulate the cell growth process using the Gillespie algorithm[75]. Each simulation ended when the population size reached 10,000 cells. Then, 1,000 cells were randomly sampled to obtain their division history.

To simulate the mutation-based cell phylogeny, we assumed that mutations randomly occur during each cell division following a Poisson distribution:

$$P(m = i) = \frac{u^i e^{-u}}{i!} \quad (8)$$

where u is the mean mutation rate per cell division. Different mutation rates (u =0.1, 0.3, or 1) were used. After obtaining the cell mutational information, we used three different algorithms to reconstruct the phylogenetic tree, respectively, namely *Maximum Likelihood* (using IQ-TREE 2[76]), *Neighbor-Joining* (using R package ape 5.6-2[77]) and *Maximum Parsimony* (using R package phangorn 2.11.1).

Simulating scRNA-seq data. We first simulated the latent expression process of genes. For each gene, we randomly generated its initial expression μ_0, drift coefficient v, and variance σ^2, and then simulated gene-specific diffusion process as follows:

$$z_{t+dt} \sim Normal(loc = z_t + vdt, scale = \sigma^2) \quad (9)$$

When cells differentiated at time t_d, for MEGs, their gene expressions remained unchanged as the same with the values in the previous process. For a non-MEG, its drift coefficient and variance were regenerated randomly and the value of expression was reset to z_{t_d}. We called the diffusion process z_t as the gene expression program. For each gene, the initial value of the expression program z_0 was randomly drawn from a gamma distribution $z_0 \sim \Gamma(0.5, 20)$, the drift coefficient v was drawn from a normal distribution $v \sim \mathrm{Normal}(0, 1)$, and the diffusion coefficient σ was drawn from a truncated normal distribution $\sigma \sim \mathrm{TruncatedNormal}\left(kz_0, \frac{|kz_0|}{3}, \min = 0.00001\right)$. In all simulations, we set the drift coefficient of

each gene expression program to change with probability 0.4 when cell differentiation happens, ultimately resulting in 10-15% of genes with unchanged expression programs upon cell differentiations, thus behaving as MEGs. The other 85-90% of genes will change dynamically with cell differentiations and thus behave as non-MEGs. Each cell was assumed to have 2,000 expressed genes, thus including 200-300 MEGs in total in each simulated dataset.

After generating the latent expression process of all genes, in order to simulate the variations introduced in real scRNA-seq experiments, the NB[78] or ZINB distribution[79] was used to obtain read/UMI count x:

$$x \sim \mathrm{ZINB}(\psi, z, \alpha), \tag{10}$$

where ψ is the zero-inflation parameter ($\psi = 1$ for negative binomial model), α is the scale parameter and the expectation of the distribution is ψz.

***Assigning the gene expression programs to the cell division history*.** Having each gene expression program $(z_1, z_2, \cdots, z_G)$ and phylogeny $\mathcal{T}$, we traversed all nodes $V \in \mathcal{T}$ and assigned the latent expression programs $z_1(d), \cdots z_G(d)$ to the nodes with branch length of d. For each gene *g*, a random number obeying the distribution $z_g(d)$ was drawn as the latent expression of that gene. By traversing all the genes and cells, the latent expression matrix can be obtained and denoted as Z. We also simulated a Gaussian noise ε to be imposed on Z, thus the latent expression matrix would be updated as $Z + \varepsilon$. Finally, given the zero-inflation factor ψ and the scale parameter α, the expression matrix X of the simulated data can be obtained by random sampling according to Equation (10).

Inference of PhyloVelo pseudotime

To infer the PhyloVelo pseudotime (forward) of each cell, we first constructed a minimum spanning tree based on the distance of cellular states on tSNE/UMAP embedding using

Prim's algorithm[80]. Thus, we can obtain a subset of the edges $\mathcal{E}$ connecting all cells together with a minimum possible total distance. We then chose any cell c_0 as the starting point and set its pseudotime $pt_{c_0} = 0$. For any other cells $c \in \cup\{e \in \mathcal{E}: c_0 \in e\}$, we calculated its pseudotime using the following equation:

$$pt_c = pt_{c_0} + \int_{\boldsymbol{x}_{c_0}}^{\boldsymbol{x}_c} \frac{1}{\boldsymbol{v}_{emb}(x)} \mathrm{d}x \tag{11}$$

where $\boldsymbol{x}_c$ is the coordinate of cell c in the embedding space and $\boldsymbol{v}_{emb}(x)$ is the phylogenetic velocity in the embedding space and varies with its coordinates.

To simplify the calculation, we replaced the velocity in this path with the average velocity of $\boldsymbol{v}(c)$ and $\boldsymbol{v}(c_0)$, denoted by $\boldsymbol{v}_a$, and used the line segments $\boldsymbol{l}_{c,c_0}$ to approximate the path. Hence, we have:

$$pt_c = pt_{c_0} + \frac{\left\|\boldsymbol{l}_{c,c_0}\right\|_2^2}{\boldsymbol{v}_a^{\mathrm{T}} \boldsymbol{l}_{c,c_0}} \tag{12}$$

Following the path generated from the minimum spanning tree, we can estimate the pseudotime of all cells and finally normalize to $[0,1]$. It should be noted that although PhyloVelo velocity fields are in backward directions, PhyloVelo pseudotime is still set to be forward as scVelo latent time.

Analysis of phylogeny-resolved scRNA-seq datasets

Datasets and pre-processing. We have applied PhyloVelo to six real phylogeny-resolved scRNA-seq datasets that are publicly available through online sources (see **Data availability**). These included *C. elegans*[14], mouse embryos[32], GEMM of lung adenocarcinoma[51], mouse xenograft models using pancreatic cancer cell line KPCY[62] and lung cancer cell line A549[63], and *in intro* culture of human kidney cell line HEK293T[64]. The embryonic lineage tree of *C. elegans* is entirely known and was obtained from http://dulab.genetics.ac.cn/TF-atlas/Cell.html, while the CRISPR-based lineage trees in

other five datasets were obtained from the original studies which were reconstructed by the mutational scars on CRISPR lineage barcodes. In the *C. elegans* dataset, because multiple synchronous embryos were pooled for the scRNA-seq experiment, many nodes in the lineage tree have been sampled multiple times. Thus, only one random cell was chosen to represent the corresponding node, while these non-repetitive cells (~300 cells) from one lineage tree constituted a "pseudo-embryo". For the mouse embryos (E8.0/8.5)[32], four (embryos 1,2,3 and 6) out of seven embryos were analyzed for their higher barcode diversity where the number of unique barcode alleles was > 500 in each embryo. For the scRNA-seq data of *C. elegans*[14] and mouse lung adenocarcinoma[51], the coordinates of tSNE or UMAP from the original studies were used. All phylogenetic trees were read and branch lengths were calculated using biopython[81] and visualized using iTOL[82]. For the scRNA-seq data of mouse embryos[32], cell lines KPCY[62], A549[63] and HEK293T[64], the dimensionality reduction and tSNE or UMAP visualization were performed using Scanpy[83] following the recommended data processing procedures and parameters as https://scanpy-tutorials.readthedocs.io/en/latest/. In each dataset, the genes with total count < 20 were filtered out.

Applying PhyloVelo. For *C. elegans*, whose embryonic cell division history is entirely known, the cell generation time was used to denote the phylogenetic distance. For the other five CRISPR/Cas9 lineage tracing datasets[32, 51, 62-64], the phylogenetic distance on a lineage tree corresponds to the number of Cas9 cutting scars on the evolving barcodes. To estimate the latent gene expressions, for *C. elegans,* the ZINB model was used to analyze the raw UMI count data because of the high-quality lineage tree. For the CRISPR/Cas9-based lineage tracing datasets, the Gaussian model was used on the post-normalized data where *normalize_per_cell()* and *log1p()* by *Scanpy*[83] were applied to the raw UMI counts. To prioritize the high-confident candidates of MEGs and speed up the computation, rather than estimating the latent expression for all genes, we firstly searched for candidate MEGs by

directly analyzing the correlation between each gene's normalized UMI counts and the phylogenetic distances to root of single cells. The top 5% of genes with the highest Spearman's correlations were first selected and then proceeded for follow-up latent expression estimations. Final MEGs were identified by the significant association (Pearson's correlation, $q<0.05$ after Benjamini-Hochberg correction; a stringent threshold $q<10^{-5}$ was used for Chan *et al*. dataset[32] given the large number of cells in individual embryos each with 6,328**-**19,071 cells) between the latent expressions and the phylogenetic distances from terminal nodes to the root of tree. The phylogenetic velocity was computed independently for each MEG. To project the phylogenetic velocity into the dimensionality reduction embedding, we built a k-nearest neighbor (kNN) graph (k=15 for *C. elegans* dataset while it was chosen by approximate to one third of total number of cells for the CRISPR lineage tracing datasets). The kNN graph was based on the Euclidean distance as the base vector and was used to estimate the coordinates of velocity embedding, as the projection of RNA velocity[22, 23].

Applying scVelo. The spliced and unspliced read counts were obtained by running velocyto (v0.6)[22] on the bam files from the output of CellRanger (6.0.2) using the raw sequence reads. To estimate RNA velocity, scVelo (version 0.2.4)[23] and the dynamical mode were used following the recommended data processing procedures as https://scvelo.readthedocs.io/VelocityBasics/. Spliced/unspliced read counts were pre-processed using the following default setting:

```
scv.pp.filter_and_normalize(adata, min_shared_counts=20, n_top_genes=2000)
scv.pp.moments(adata, n_neighbors=30, n_pcs=30)
```

Applying VeloVAE. VeloVAE[58] applies the same data preprocessing steps as scVelo. There are three data training models including Basic Model (assuming fixed transcription rates), Full Model (assuming variable transcription rates) and Full VB Model (treating the

rate parameters as random variables). Full Model was used as recommended in the paper. The model training parameters were used following the example in its GitHub repository (https://github.com/welch-lab/VeloVAE/blob/master/notebooks/velovae_example.ipynb).

```
vae = vv.VAE(adata, tmax=20, dim_z=5)
vae.train(adata, gene_plot=gene_plot, plot=True, figure_path=figure_path)
```

Applying DeepVelo. DeepVelo[59] also applies the same data preprocessing steps as scVelo. Model configurations were the same as the default setting (https://github.com/bowang-lab/DeepVelo/blob/main/examples/figure2.ipynb) except some updates as following:

```
configs = dict(
    "name": "DeepVelo", # name of the experiment
    "loss": dict("args": dict("coeff_s": autoset_coeff_s(adata))),
    "trainer": dict("verbosity": 0), # increase verbosity to show training progress
    "n_gpu":0
)
configs = update_dict(Constants.default_configs, configs)
```

Applying cellDancer. cellDancer[60] applies the same data preprocessing steps as scVelo. The format conversion of the data is according to its tutorial (https://guangyuwanglab2021.github.io/cellDancer_website/index.html), and the velocity inference uses all genes and proceeds according to the default parameters. The velocity field is visualized using the Dynamo[50].

Applying UniTVelo. UniTVelo[61] also applies the same data preprocessing steps as scVelo. Model configurations were the same as the default setting (https://unitvelo.readthedocs.io/en/latest/Figure2_ErythroidMouse.html):

```
velo_config = utv.config.Configuration()
velo_config.R2_ADJUST = True
velo_config.IROOT = None
velo_config.FIT_OPTION = '1'
velo_config.AGENES_R2 = 1
```

Applying Dynamo. Dynamo[50] was used to infer the quantitative cell-state transition matrix and visualize cell state transition graph. We use the velocity field inferred by PhyloVelo as input and calculate the transition matrix as follows:

```
dyn.vf.VectorField(adata, basis='umap', M=1000, pot_curl_div=True)
dyn.vf.topography(adata, basis='umap')
dyn.ext.ddhodge(adata, basis='umap')
dyn.pd.state_graph(adata, group='cell_states', basis='umap', method='vf', approx=False)
```

Analysis of static barcoding-based lineage tracing datasets

Datasets and pre-processing. We have applied the extended model of PhyloVelo to two static barcoding datasets including LARRY hematopoietic differentiation[37] and intratumoral CD8+ T cells in BCC[57] that are publicly available through online sources (see **Data availability**). For the LARRY dataset, lentiviral barcoding data at day 6 was used to obtain the clone size information for each cell. For the CD8+ T cells data, the TCR specificity clones were identified by GLIPH[84] which defines clones based on the following two criteria: 1) global similarity, TCR sequences within the same T cell clone have at most one amino acid difference; 2) local similarity, two TCRs in same clone contain an identical CDR3 motif, which is 2-4 k-mer amino acids in length and is significantly enriched from random sub-sampling of unselected repertoires. To avoid batch effect, patient 9 with the largest cell number (4,659 cells) was selected for identification of MEGs and inference of phylogenetic velocities. The phylogenetic velocities were then transferred to all CD8+ T cells (12,788

cells) from all 12 BCC patients. Cells whose clonal barcodes were not determined were filtered out. The coordinates of dimensionality reduction embedding, SPRING (LARRY dataset) or UMAP (T cell dataset), from the original studies were used for visualization. In each dataset, the genes with total count < 20 were filtered out.

Applying PhyloVelo. For both datasets, the phylogenetic time of a cell corresponded to the logarithm of the clone size. To estimate the latent gene expressions, Gaussian process model was used on the post-normalized data. Same as the scRNA-seq data analysis above, *normalize_per_cell()* and *log1p()* by *Scanpy*[83] were applied to the raw UMI counts. The top 5% genes with the highest Spearman's correlation between normalized gene expression and phylogenetic time were first selected, then proceeded for follow-up latent expression estimations. Final MEGs were identified by the significant association (Pearson's correlation, q<0.05) between the latent expressions and the logarithm of clone size. Projecting phylogenetic velocities into the embedding followed the same procedure as CRISPR lineage tracing data analysis.

Transferring the phylogenetic velocities of MEGs to independent datasets

To evaluate whether the phylogenetic velocities of MEGs estimated from one phylogeny-resolved scRNA-seq dataset are sufficiently robust to infer the velocity fields in independent datasets in the absence of phylogenetic information, three datasets were analyzed including *C. elegans*[14]*,* mouse erythroid development[19, 32], and the GEMM of lung adenocarcinoma[51]. Here, the MEGs and corresponding phylogenetic velocity estimates were directly applied to another scRNA-seq datasets in similar biological conditions. For *C. elegans,* we applied the phylogenetic velocities from AB lineage in a single pseudo-embryo (n=298 cells) to all AB lineage cells (n=29,600) in multiple embryos. We also applied them to non-AB lineages that differentiate to hypodermis, body wall muscles (BWM) and pharynx, respectively. For mouse erythroid differentiation, we applied the phylogenetic velocity estimates in the

erythroid lineage cells from a single embryo (E8.5, n=2,419 cells) of the Chan *et al*. dataset[32] to the other three embryo and the temporally-sequenced mouse embryos (E6.5-E8.5, n=12,324 cells) of the Pijuan-Sala *et al*. dataset[19]. We also applied the phylogenetic velocities of LT-MEGs identified from the Chan *et al*. dataset[32] to predict the entire embryo development with the Pijuan-Sala *et al*. dataset[19] (E6.5-E8.5, n=10,000 out of 116,312 cells were randomly sampled), and predict mouse brain development with Manno *et al*. dataset[48] (E7-18, n=10,000 out of 292,495 cells were randomly sampled). For lung adenocarcinoma[51], the phylogenetic velocity estimates in one KP primary lung tumor (3726_NT_T1, n=754 cells) were applied to all 58,022 single cells from all pooled KP primary lung tumors. Finally, for intratumoral CD8+ T cells, in order to avoid the batch effect, the phylogenetic velocity estimates in 4,659 cells from patient 9 were applied to all 12,788 CD8+ T cells from 12 BCC patients.

Gene ontology (GO) enrichment analysis

GO enrichment analysis was performed using clusterProfiler v4.4.4[85]. The cutoff for *p* value and *q* value were set to 0.05 and 0.25, respectively. After excluding Cellular Components (CC) terms, all significant terms were retained for downstream analyses. Subsequently, top 20 GO terms of each sample were merged and these terms were sorted by their total occurrence and mean *q* value across samples. Finally, the top 20 GO terms enriched were visualized using ggplot2 v3.4.0.

Data availability

All data analyzed in this article are publicly available through online sources. The annotated data, lineage trees, results and Python implementation are available at https://phylovelo.readthedocs.io/. The raw data for the *C. elegans* dataset[14] can be accessed with GSE126954 and the lineage tree can be accessed from http://dulab.genetics.ac.cn/TF-atlas/Cell.html. The CRISPR lineage tracing datasets from

the mouse embryos[32, 86] can be accessed with GSE117542. The single cell RNA-seq data of mouse brain development [48] can be accessed with PRJNA637987. The time-course single-cell RNA-seq data of whole mouse embryos (E6.5-8.5)[19] can be accessed with E-MTAB-6967. The dataset of mouse primary lung tumors[51] can be accessed with PRJNA803321 and from Zenodo (https://zenodo.org/record/5847462#.Yt4-PewRXUI). The dataset of mouse pancreatic cancer cell line KPCY[62] can be accessed with GSE173958 and from Mendeley (https://doi.org/10.17632/t98pjcd7t6.1). The dataset of human lung cancer cell line A549[63] can be accessed with GSE161363. The dataset of human kidney cell line HEK293T[64] can be accessed with PRJNA757179. The LARRY lentiviral barcoding dataset of hematopoiesis[37] can be accessed with GSE140802. The single-cell TCR and RNA sequencing data of T cells in BCC[57] can be accessed with GSE123813.

Code availability

PhyloVelo[87] is freely available as Python package at https://github.com/kunwang34/PhyloVelo. Detailed workflows to reproduce figures and results in this paper are written as Jupyter notebook in the repository. The annotated data, lineage trees, results and Python implementation are available at https://phylovelo.readthedocs.io/.

Methods-only references

75. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry* **81**, 2340-2361 (1977).

76. Minh, B.Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).

77. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

78. Chen, W. et al. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol* **19**, 70 (2018).

79. Jia, C. Kinetic foundation of the zero-inflated negative binomial model for single-cell RNA sequencing data. *SIAM Journal on Applied Mathematics* **80**, 1336-1355 (2020).

80. Prim, R.C. Shortest connection networks and some generalizations. *The Bell System Technical Journal* **36**, 1389-1401 (1957).

81. Cock, P.J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).

82. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021).

83. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

84. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94-98 (2017).

85. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

86. Chan, M.M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82 (2019).

87. Wang, K., Hou, L., Wang, X., Zhai, X., Lu, Z., Zi, Z., Zhai, W., He, X., Curtis, C., Zhou, D., & Hu, Z. PhyloVelo, Phylogeny-based transcriptomic velocity of single cells. GitHub. https://github.com/kunwang34/PhyloVelo (2023).

Figures

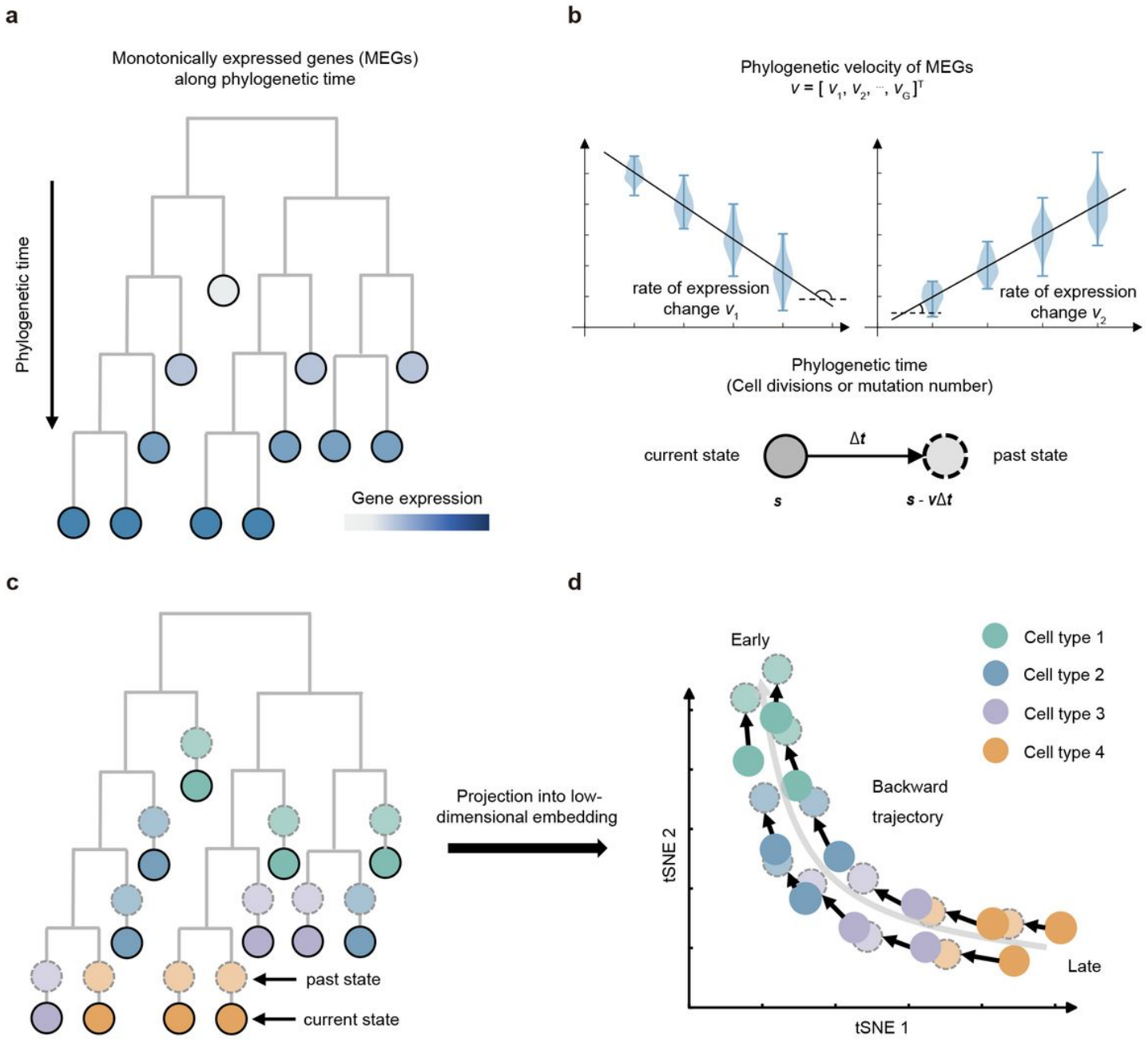

Figure 1

Schematic of the PhyloVelo framework. (**a**) Schematic of monotonically expressed genes (MEGs) over phylogenetic time on a cell phylogenetic tree. (**b**) Two examples of MEGs whose latent expressions are associated with the phylogenetic time (cell divisions or mutation number). A diffusion process of gene expressions was used to model the changes of latent expressions over phylogenetic time. This enables the estimation of the phylogenetic velocity, $= (\ _1, \ _2, \ldots, \ _G)$, which corresponds to the drift coefficients of G MEGs in the diffusion process (approximate to the slope of linear regression between latent expression and phylogenetic time). Whiskers: minimum and maximum. (**c**) Phylogenetic velocity predicts the past transcriptional state of a cell before a unit of phylogenetic time (one cell division or mutation). (**d**)

Projection of the phylogenetic velocity into low dimensional embedding enables the mapping of cell-state trajectory in backward directions.

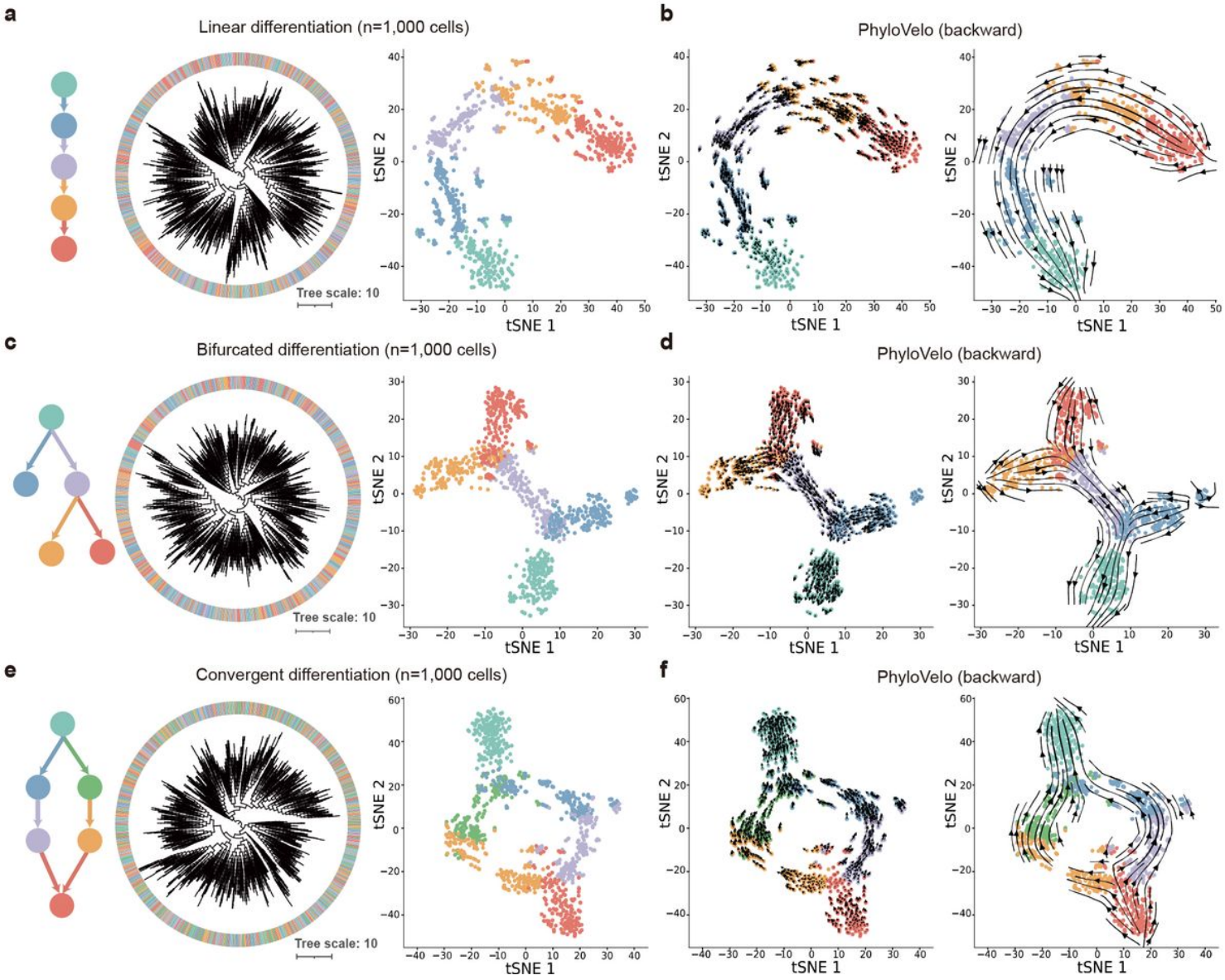

Figure 2

PhyloVelo recovers complex cell lineages in simulations. Simulation of single-cell RNA-seq data and paired cell-division history under linear (**a**), bifurcated (**b**), and convergent (**c**) differentiation models, respectively. Colors are labeled by cell types. Each simulation consists of 1,000 cells randomly sampled from a growing cell population at 10,000 cells. Each cell has 2,000 expressed genes, including 200-300 MEGs. (**d-f**) Phylogenetic velocity fields reconstructed by PhyloVelo for the corresponding differentiation scenarios. The left panel shows the single-cell level of velocity fields, while the right panel shows the same velocity fields visualized as streamlines in scVelo. PhyloVelo velocity fields are at backward directions.

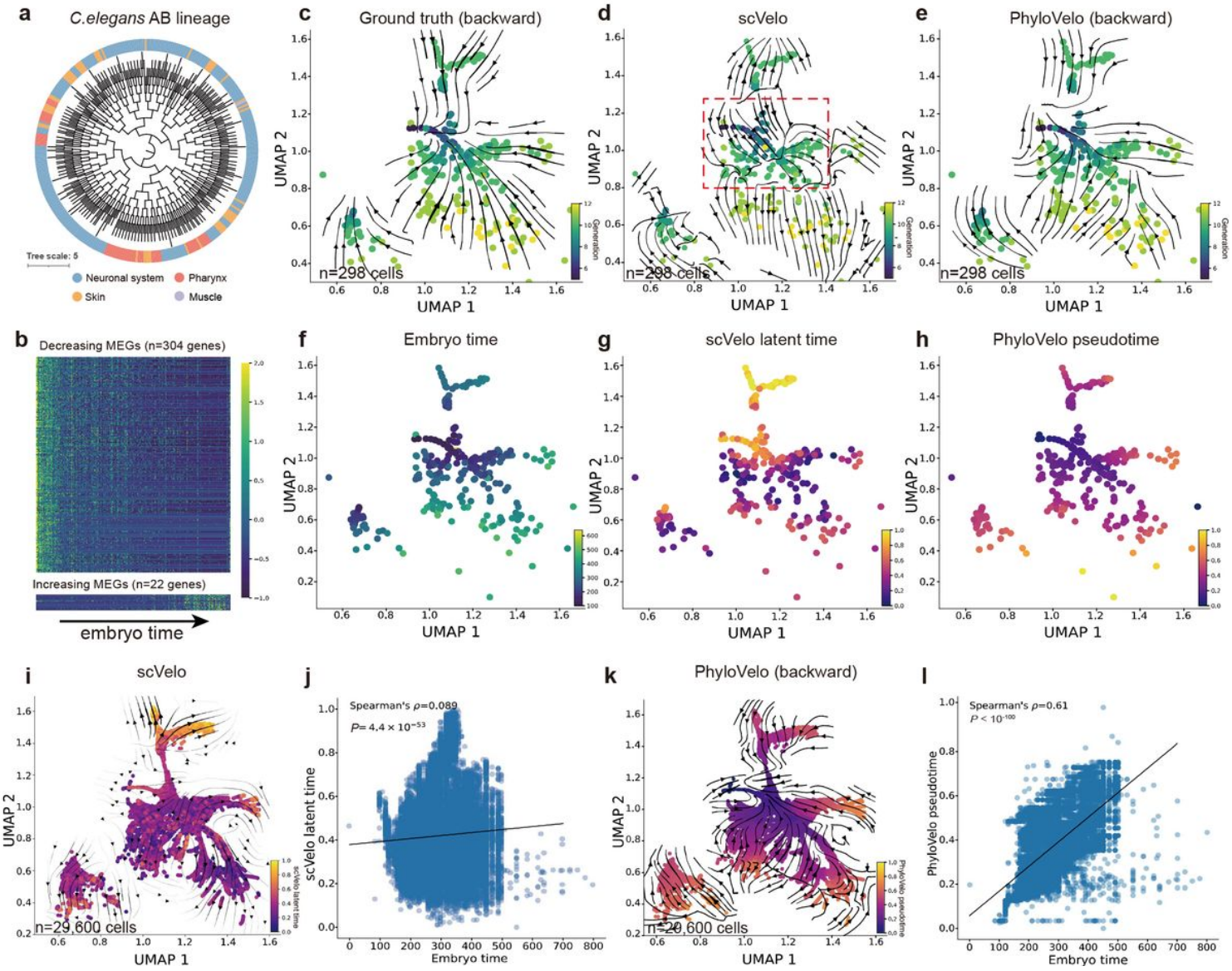

Figure 3

PhyloVelo reconstructs the embryonic differentiation trajectories of *C. elegans*. (**a**) Phylogenetic tree of the *C. elegans* AB lineage. (**b**) Heatmap showing the expressions (z-score normalized) of MEGs along *C. elegans* embryo time. (**c**) The ground-truth velocity fields represent vectors superimposed on the cells that point to their immediate parental cells on the Uniform Manifold Approximation and Projection (UMAP) plot. (**d-e**) The velocity fields estimated by scVelo (dynamical mode) (**d**) or PhyloVelo (**e**). Dash square indicates the early embryonic lineages where RNA velocity gave erroneous estimations on the fate directions. (**f**) *C. elegans* embryo time as Packer *et al.*[14]. (**g**) scVelo latent time. (**h**) PhyloVelo pseudotime. (**i**) RNA velocity fields for all 29,600 AB lineage cells. Colors are labeled by scVelo latent time. (**j**) The correlation between scVelo latent time and embryo time for all AB lineage cells. (**k**) PhyloVelo velocity fields for all 29,600 AB lineage cells, estimated by the phylogenetic velocity of MEGs in a single embryo (n=298 cells). Cell colors are labelled by PhyloVelo pseudotime. (**l**) The correlation between PhyloVelo pseudotime and embryo time for all AB lineage cells. The Spearman correlation coefficients and *P* values are shown.

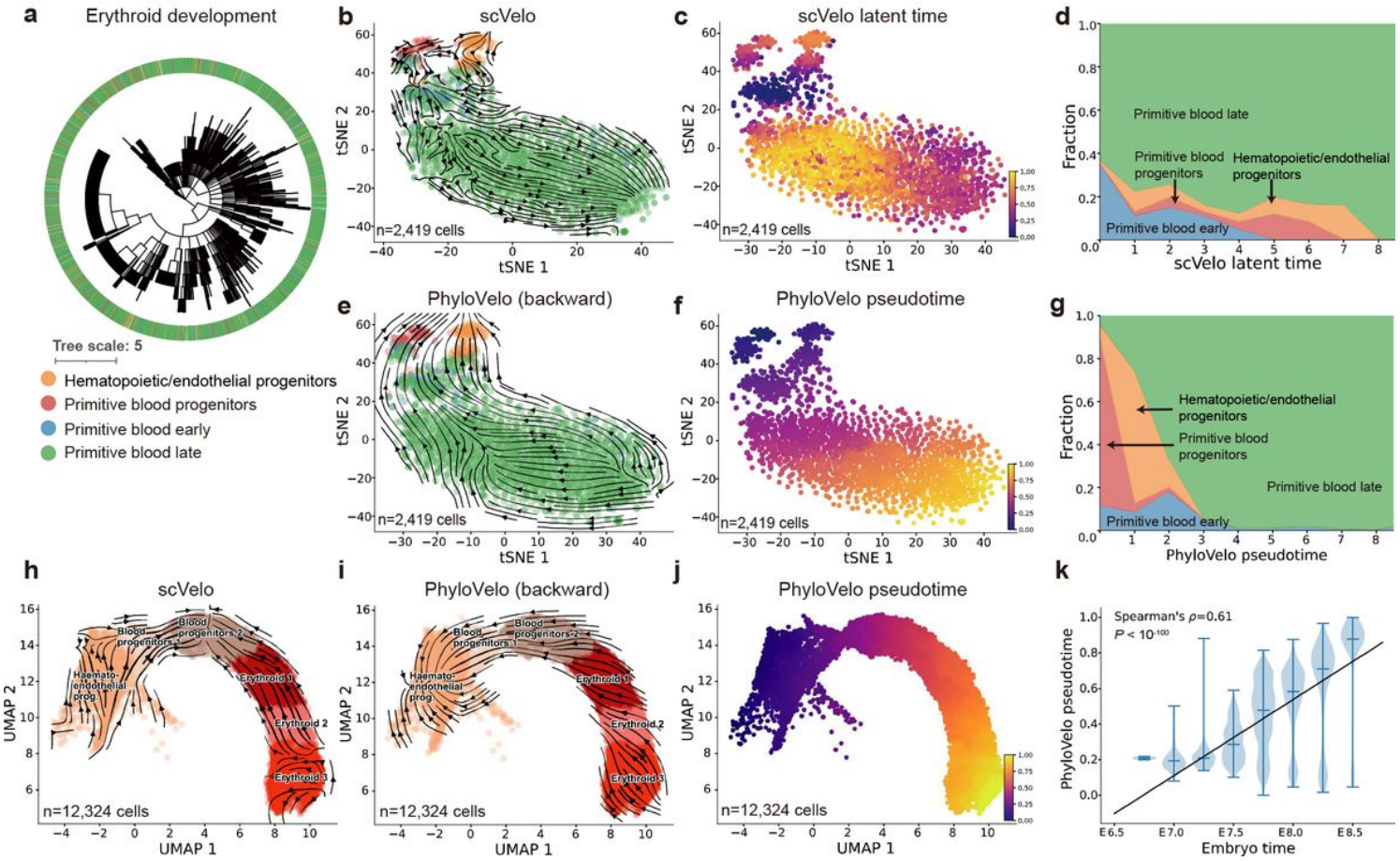

Figure 4

PhyloVelo reconstructs the cellular trajectory of mouse erythroid maturation. (**a**) Phylogenetic tree of the 2,419 erythroid lineage cells (embryo 3, E8.5) in Chan *et al*. dataset[32]. (**b-c**) RNA velocity fields (scVelo - dynamical mode) and the latent time of mouse erythroid development. (**d**) Muller plot showing the fractions of four cell types that change over scVelo latent time. (**e-f**) PhyloVelo velocity fields and the pseudotime of mouse erythroid development. (**g**) Muller plot showing the fractions of four cell types that change over PhyloVelo pseudotime. (**h**) Erroneous estimations of RNA velocity fields on erythroid maturation because of multiple rate kinetics (MURK). Data were from Pijuan-Sala *et al*.[19]. (**i**) PhyloVelo velocity fields of erythroid maturation for Pijuan-Sala *et al*. dataset while using the MEGs identified from Chan *et al*. dataset. (**j**) PhyloVelo pseudotime of erythroid maturation in Pijuan-Sala *et al*. dataset. (**k**) The correlation between PhyloVelo pseudotime and mouse embryo time (n=12,324 cells). The Spearman correlation coefficient and *P* value are shown here. Whiskers: minimum and maximum; center lines: median.

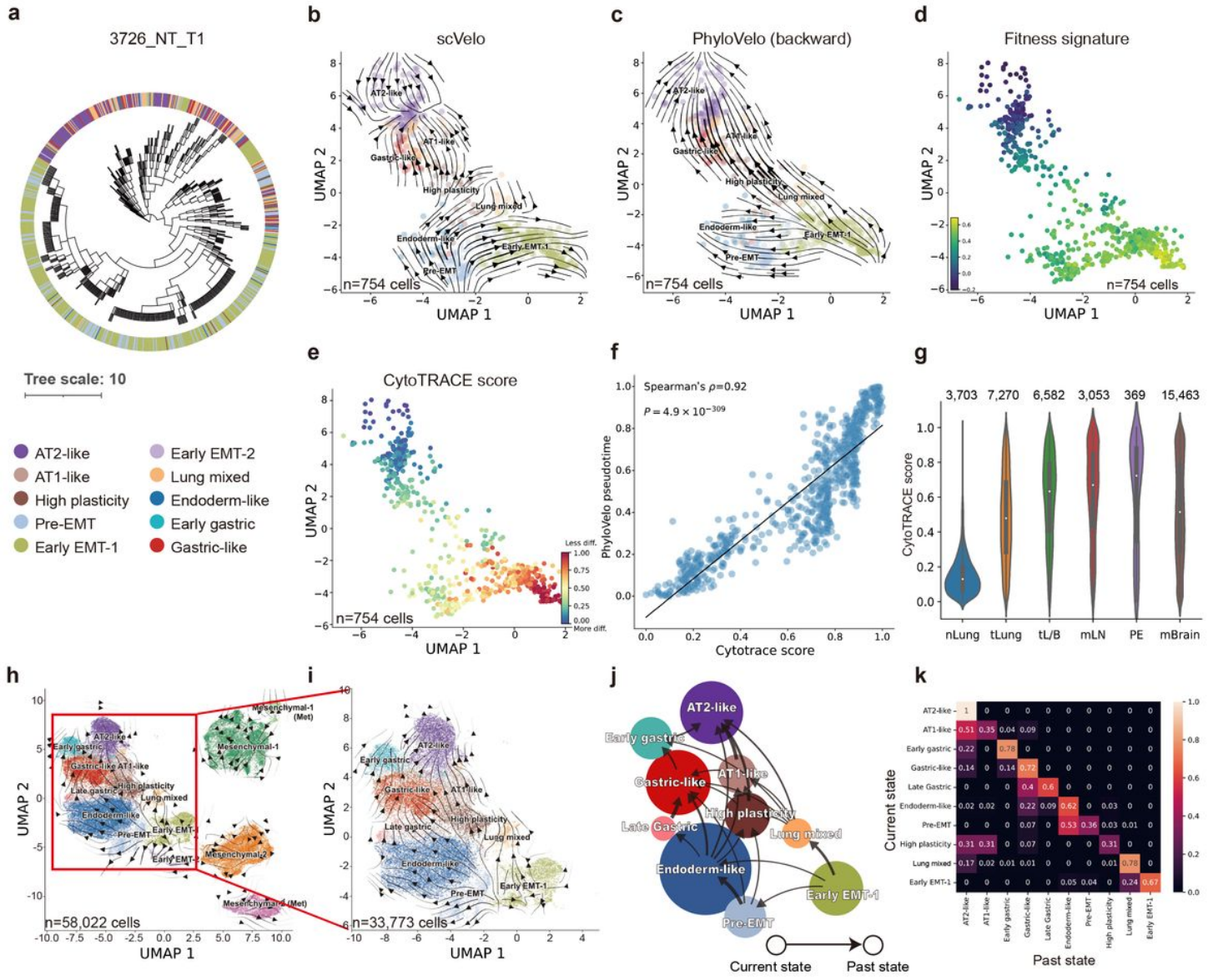

Figure 5

PhyloVelo identifies a dedifferentiation trajectory in lung tumor evolution. (**a**) Phylogenetic tree of 754 cells from a KP-mouse primary lung tumor, 3726_NT_T1, in Yang *et al*. dataset[51]. The scRNA-seq data, cell type annotations, and lineage trees were obtained from the original study. (**b**) RNA velocity fields (scVelo - dynamical mode). (**c**) PhyloVelo velocity fields. (**d**) Fitness signatures of individual cells, as defined by Yang *et al*. (**e**) CytoTRACE score of individual cells. (**f**) The correlation between PhyloVelo pseudotime and CytoTRACE scores. The Spearman correlation coefficient and *P* value are shown here. (**g**) CytoTRACE score of single tumor cells from human lung primary sites (tLung and tL/B), pleural fluids (PE), lymph node metastases (mLN), and brain metastases (mBrain), as well as normal tissues from lungs (nLung), as described in Kim *et al*.[54]. Bar, median; box, 25th to 75th percentile (IQR); vertical line, data within 1.5 times the IQR. (**h**) PhyloVelo velocity fields for all 58,022 single cells from pooled KP primary lung tumors, estimated by the MEGs identified from 3726_NT_T1. (**i**) PhyloVelo velocity fields for the cell types that existed in 3726_NT_T1. (**j**) Cell-type transition graph (backward) based on the transition

rate matrix between any two cell types (**k**), estimated by Dynamo using PhyloVelo velocity fields as input. The arrows point from the current states to the past states.

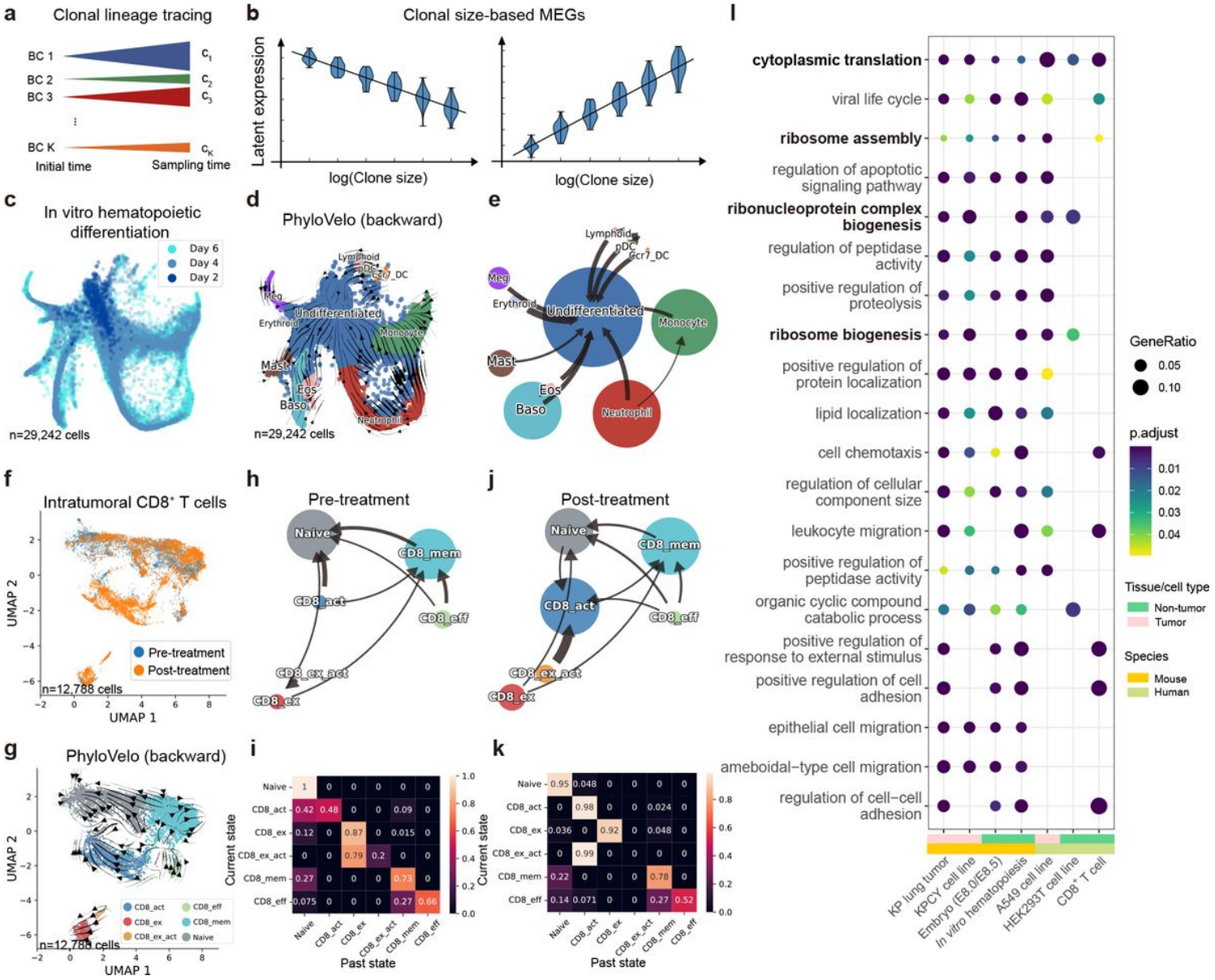

Figure 6

PhyloVelo inference with clonal lineage tracing data and MEGs are enriched in ribosome-mediated processes. (**a**) Schematic of clonal lineage tracing data where static barcodes identify cells of common ancestry. Clone size, denoted by c_k for k clones, represents the number of cells carrying the same unique barcode. (**b**) Two examples of clonal size-based MEGs whose latent expressions are positively or negatively associated with the logarithm of clone sizes, respectively. Whiskers: minimum and maximum. (**c**) scRNA-seq data of in vitro hematopoietic differentiation from Weinreb *et al.*[37], where each cell over the course of 2, 4, and 6 days culture could be traced by one unique barcode. (**d**) The velocity fields estimated by PhyloVelo. (**e**) Cell type transition graph (backward) of in vitro hematopoietic differentiation. (**f**) UMAP of tumor-infiltrating CD8+ T cells in BCC samples pre- and post-PD-1 blockade, colored by anti-PD-1

treatment status. Data were from Yost *et al*.[57] (**g**) The velocity fields estimated by PhyloVelo. (**h-i**) Cell-type transition graph and transition matrix (backward) at pre-treatment. (**j-k**) Cell-type transition graph and transition matrix (backward) at post-treatment. CD8_act: CD8+ activated T cells; CD8_ex: CD8+ exhausted T cells; CD8_ex_act: CD8+ exhausted/activated T cells; CD8_eff: CD8+ effector T cells; CD8_mem: CD8+ memory T cells. (**l**) Gene ontology (GO) enrichment of MEGs identified across tissues and organisms. The top and most commonly shared 20 biological processes are shown. Ribosome-mediated processed are highlighted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ReportingSummary0622.pdf
- SupplementaryInformation0622.pdf
- SupplementaryTable1.xlsx
- ExtendedDataFig.1.jpg
- ExtendedDataFig.2.jpg
- ExtendedDataFig.3.jpg
- ExtendedDataFig.4.jpg
- ExtendedDataFig.5.jpg
- ExtendedDataFig.6.jpg
- ExtendedDataFig.7.jpg
- ExtendedDataFig.8.jpg
- ExtendedDataFig.9.jpg
- ExtendedDataFig.10.jpg