Bioinformatics, ???, 1–7
doi: ???
???
???

Assessing RNA velocity stability across synthetic replicates using count splitting

Yuhong Li,[1] Zeyu Jerry Wei,[2] Yen-Chi Chen[2] and Kevin Z. Lin[1,*]

[1]Department of Biostatistics, University of Washington, 3980 15th Avenue NE, 98195, WA, USA and [2]Department of Statistics, University of Washington, 4110 E Stevens Way NE, 98195, WA, USA

*Corresponding author. kzlin@uw.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

RNA velocity is a computational framework that enables predicting future cell states based on single-cell RNA sequencing data and offers valuable insights into dynamic biological processes. However, there is a lack of general methods to quantify the uncertainty and stability of these predictions from various RNA velocity methods. In this paper, we present a novel framework for evaluating the stability of RNA velocity estimates by generating two independent splits of the original data using negative binomial count splitting. We apply our workflow to five RNA velocity methods and two well-characterized datasets about mouse erythroid development and pancreatic endocrinogenesis. Our results highlight key differences between methods, identify inconsistencies such as reversed velocity flow between different RNA velocity methods, and demonstrate robustness when certain intermediary cell states are missing. Our framework is broadly applicable across RNA velocity methods and provides a robust way to assess and compare method performance. This generalizability makes our approach an important tool for selecting reliable RNA velocity methods in diverse biological contexts.

Key words: Single-cell RNA-seq, Developmental biology, Negative Binomial, Trajectory analysis, Statistical inference

1. Introduction

RNA velocity is a widely used computational tool in single-cell RNA sequencing (scRNA-seq) analyses that provides new insights into temporal transcriptomic dynamics [La Manno et al., 2018]. By modeling the rate of change of the number of spliced and unspliced mRNA transcripts via an ordinary differential equation (ODE), these methods estimate high-dimensional velocity vectors for individual cells, capturing the direction of their developmental trajectories. When projected onto a 2D embedding, these velocity vectors enable the visualization of the entire cell differentiation process. The RNA velocity framework has been paramount in making critical discoveries in understanding the precise order of intermediary cell states in systems such as erythroid maturation [Barile et al., 2021], muscle self-renewal [Hettinger et al., 2023], and embryogenesis [Qiu et al., 2022]. This importance has fueled the development of many follow-up RNA velocity methods, each making different assumptions and statistical models. Although these RNA velocity methods exist to predict cell development, each can yield different results, making it difficult to assess their reliability and the uncertainty inherent in their predictions [Marot-Lassauzaie et al., 2022, Gorin et al., 2022, Bergen et al., 2021, Ancheta et al., 2024, Zheng et al., 2023].

There are already metrics to quantify the RNA velocity uncertainty, but these metrics are often specific to certain RNA velocity estimators. For example, VeloVI [Gayoso et al., 2024] uses variational inference and introduces intrinsic and extrinsic uncertainty computations. The intrinsic uncertainty is measured using samples from the posterior distribution by calculating the cosine similarity between each sampled velocity vector and the posterior mean velocity vector for each cell. Extrinsic uncertainty, in contrast, assesses the variance in cosine similarities between a cell's sampled velocity vectors and those of its nearest neighbor. Although these measures quantify uncertainty in velocity predictions, they are specific to VeloVI's Bayesian framework and are not generalizable to other RNA velocity methods.

In addition, scVelo [Bergen et al., 2020] implements a velocity confidence metric that measures the coherence of velocity vector estimates within neighborhoods. By computing the mean correlation between each cell's velocity vector and its neighbors', the score measures the smoothness of the velocity vector field. While this metric quantifies the local agreement of velocity vectors, it does not adequately evaluate the uncertainty of an RNA velocity method. Since neighboring cells tend to have similar transcriptomic profiles, they often exhibit similar velocity vectors,

© The Author ???. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

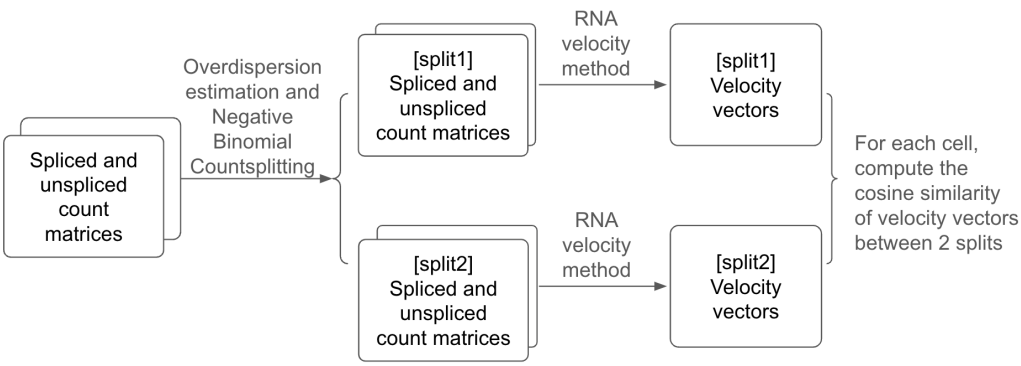

Fig. 1: Diagram illustrating our workflow. Starting with the initial spliced and unspliced count matrices, we estimate the overdispersion parameter and perform negative binomial count splitting on both matrices to obtain two sets of spliced and unspliced datasets. We then run an RNA velocity method on each split separately and get two sets of cell-specific velocity vectors. The replicate coherence for each cell is computed by the cosine similarity of its two RNA velocity vectors, one from each split.

resulting in high confidence scores approaching one. Although such scores indicate local smoothness and coherence, they do not truly reflect the uncertainty inherent in the method's output [Ancheta et al., 2024]. This motivates us to devise a framework that captures the uncertainty in RNA velocity methods, which is not confounded by the local similarity of cells within a neighborhood.

Negative binomial count splitting, developed by [Neufeld et al., 2023], is a method that decomposes any $n \times p$ matrix that follows a negative binomial distribution into multiple independent datasets. This idea has been the focus of many recent statistical theory works [Leiner et al., 2023, Dharamshi et al., 2024], and has been used to the clustering stability of scRNA-seq data [Neufeld et al., 2024]. Count splitting aims to create multiple $n \times p$ matrices (typically two) that entry-wise sum to the original matrix while ensuring that each split matrix is independent of the others. The resulting splits can be interpreted as synthetic "technical replicates." Because the sequencing counts for any particular cell are statistically independent across the splits, we can interpret count splitting as repeating a sequencing experiment on the exact same set of cells multiple times. In reality, it is only possible to sequence the same cell once, and these technical replicates are often inaccessible due to practical constraints. Nonetheless, we can synthetically generate independent datasets from the same biological process using count splitting. This statistical framework allows us to compare the behavior of the same cells and quantify the stability of different RNA velocity methods across independent replicates. In this work, we introduce this paradigm incorporating count splitting to quantify the uncertainty in different RNA velocity methods. We will refer to velocity confidence defined in Bergen et al. [2020] as *local coherence* and our uncertainty measure as *replicate coherence* in the following context.

2. Materials and methods

We present a framework for evaluating the performance of different RNA velocity methods by creating two independent splits of the scRNA-seq data through negative binomial count splitting [Neufeld et al., 2023] and comparing RNA velocity estimates between these splits (Fig. 1).

Given a dataset with n cells and p genes, we denote the data matrix as $\mathbf{X} \in \mathbb{Z}_{\geq 0}^{n \times p}$. Each entry $\mathbf{X}_{ij}$ is a non-negative integer representing the expression count of the j-th gene in cell i. Assume that $\mathbf{X}_{ij} \sim \mathrm{NB}(\mu_{ij}, b_j)$, a negative binomial with mean $\mathbb{E}[\mathbf{X}_{ij}] = \mu_{ij}$ and variance $Var(\mathbf{X}_{ij}) = \mu_{ij} + \mu_{ij}^2/b_j$, where each $b_j > 0$ denotes a gene-specific overdispersion parameter. A finite non-negative b_j indicates overdispersion where the variance exceeds the mean, and an infinite b_j indicates the distribution approaches Poisson(μ_{ij}). For each gene, we fit a Gamma-Poisson Generalized Linear Model [Ahlmann-Eltze and Huber, 2020] to estimate its overdispersion parameter as an input for the count splitting procedure. In our analysis, each dataset consists of a spliced and an unspliced data matrix, and we estimate the overdispersion parameters for spliced and unspliced matrices separately.

After estimating the overdispersion parameters, we apply negative binomial count splitting to generate two splits $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ with an equal probability of 0.5. For each entry,

$$(\mathbf{X}_{ij}^{(1)}, \mathbf{X}_{ij}^{(2)}) \mid \mathbf{X}_{ij} = x_{ij} \sim \mathrm{DirichletMultinomial}(x_{ij}, \frac{1}{2}b'_j, \frac{1}{2}b'_j),$$

where b'_j is the estimated overdispersion parameter for gene j. The theoretical properties of count splitting proved in Neufeld et al. [2023] show that: 1) the splits $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are conditionally independent given $\mathbf{X}$, and 2) both $\mathbf{X}_{ij}^{(1)}, \mathbf{X}_{ij}^{(2)} \sim \mathrm{NB}(\frac{1}{2}x_{ij}, \frac{1}{2}b'_j)$ for every cell i and gene j.

After count splitting, we have two independent count matrices that sum to the original counts, and each preserves a negative binomial distribution. Treating the two datasets as technical replicates of the same set of cells, we run the same RNA velocity method on each separately. We review the basics of RNA velocity in Supplementary Information S1. In brief, RNA velocity models the cell dynamics by positing an ordinary differential equation that relates the number of spliced to unspliced mRNA transcripts for each gene. For RNA velocity methods that explicitly require separate count matrices for the spliced and unspliced counts, we separately apply count splitting to the observed spliced and unspliced counts. In this case, we end up with two spliced count matrices and two unspliced count matrices. After applying an RNA velocity method on each split (i.e., a pair of spliced and unspliced matrices), we get two sets of cell-specific velocity

matrices $\mathbf{V}^{(1)}, \mathbf{V}^{(2)} \in \mathbb{R}^{n \times p}$ for $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ respectively, where i-th row in $\mathbf{V}^{(s)}$, denoted as $v_i^{(s)}$ is the velocity vector for cell i across all p genes based on the data in the s-th split. To assess the stability of the results, we compare the cell-specific velocity vectors from the two splitted datasets and measure the alignment per cell using cosine similarity,

$$\frac{(v_i^{(1)})^\top v_i^{(2)}}{||v_i^{(1)}||_2 ||v_i^{(2)}||_2}$$

for each cell i where $\|\cdot\|_2$ denotes the Euclidean norm. A cosine similarity value close to $+1$ indicates that the two velocity vectors for a specific cell are similar in direction, suggesting stability in the velocity estimate, and a value close to -1 reflects opposite directions between the two velocity vectors and divergence in the results.

This framework offers the advantage of being method-agnostic, as it is not limited to a specific RNA velocity approach and can be applied to any method of interest. Additionally, splitting the data into two subsets with identical cell sets enables pairwise comparisons of the velocity estimates for each cell. In contrast, scVelo's local coherence compares each cell's velocity estimate to its neighbors' under some centering operations. Specifically, let v_i denote the velocity vector of cell i, $\mathrm{ne}(i)$ denote the set of neighboring cells of cell i, and p denote the number of genes in each velocity vector. Further, let v_i' denote the centered velocity vector for cell i, where each element v_{ij}' is defined as $v_{ij}' = v_{ij} - \frac{1}{p}\sum_{j=1}^{p} v_{ij}$. The local coherence for cell i is then computed as

$$\frac{1}{|\mathrm{ne}(i)|} \sum_{k \in \mathrm{ne}(i)} \frac{v_k'^\top v_i'}{||v_k'||_2 ||v_i'||_2},$$

which is essentially the mean cosine similarity between a cell's centered velocity vector and those of its neighbors, capturing the local smoothness. While both metrics provide measures of method performance from different perspectives, it is important to note that local coherence relies on information from the pre-computed neighborhoods derived from kernel smoothing. This smoothing step can artificially inflate local coherence and potentially limit its ability to fully reflect true performance. However, replicate coherence preserves the comparability of two data splits during velocity computation, providing a clearer assessment of method stability.

We apply our workflow to five RNA velocity methods: scVelo [Bergen et al., 2020], UniTVelo [Gao et al., 2022], scTour [Li, 2023], the standard VeloVI with the preprocessing step, and VeloVI without the preprocessing step [Gayoso et al., 2024]. scVelo and VeloVI both model spliced and unspliced mRNA counts, with scVelo estimating gene-specific kinetics and VeloVI providing uncertainty quantification. UniTVelo takes a spliced-only approach with a unified latent time across genes, while scTour uniquely models total RNA counts, projecting velocities to a low-dimensional space for visualization. We evaluate all RNA velocity five methods on two scRNA-seq datasets, including a mouse erythroids during gastrulation [Pijuan-Sala et al., 2019] and a mouse pancreatic endocrinogenesis development [Bastidas-Ponce et al., 2019] using both our replicate coherence and scVelo's local coherence. The details of the datasets are provided in Supplementary Information S2 and Supplementary Table S1. In Supplementary Information S3, we document how we specifically compute the replicate coherence for each method since some methods differ in what they take as input or produce as output. Importantly, our method requires computing the correlation between the velocity vectors estimated from each split. However, methods like scTour instead output the RNA velocity in a latent embedding space. In Supplementary Information S3, we document how to handle scenarios like this and demonstrate the validity of our procedure in Supplementary Figure S8.

3. Results

3.1. Evaluation of replicate coherence in RNA velocity analyses

We use the mouse erythroid dataset to demonstrate the insights our replicate coherence brings to an RNA velocity analysis. This dataset contains 9815 cells and 53,801 genes, with the true developmental order being well-studied where the blood progenitor cells gradually differentiate into erythroid cells [Pijuan-Sala et al., 2019]. Using this dataset, we show that our replicate coherence can aid in assessing the robustness of the results from an RNA velocity analysis. Our premise is that if we apply count splitting and apply the same RNA velocity workflow on each split, a misalignment between the two resulting RNA velocity fields indicates potential computational instability in the method.

We first validate the theoretical result of independence among genes between the two splits. Among genes with nonzero counts in both splits, most exhibit correlations close to 0, supporting the notion of treating the two splits as technical replicates (Fig. 2a). We provide additional investigations in Supplementary Figure S1, where we visualize the impact of count splitting in more detail and ensure the overdispersion estimates are reliable by assessing how they relate to the sparsity of the genes.

Next, running all five methods on the dataset, we find UniTVelo, scTour, and VeloVI without the preprocessing step can correctly identify the trajectory. In contrast, scVelo and standard VeloVI produce reversed predictions, with velocity flow pointing from erythroid cells back to blood progenitor cells (Fig. 2b). Such a pattern is not directly reflected in the measure of local coherence, where all methods except the standard VeloVI show high values close to 1, indicating the overall smoothness of velocity vectors within cell neighborhoods (Fig. 2c). However, regarding replicate coherence, scVelo and standard VeloVI exhibit median values as low as 0.5, suggesting relative instability across technical replicates and corresponding to less precise performance. In contrast, the other methods' replicate coherence values stay relatively high. We reiterate that our replicate coherence was computed directly based on the velocity vectors $v_i^{(s)} \in \mathbb{R}^p$ for each cell i between splits $s = 1$ and $s = 2$, i.e., the rate of change in the spliced counts across all the genes in the analysis. We emphasize this since work like Zheng et al. [2023] has emphasized the ambiguity of projecting RNA velocity vectors onto a UMAP embedding. The authors showed that the UMAP visualizations, such as those in Figure 2b, could sometimes be misleading. However, since our replicate coherence is computed based on the velocity vectors across all the genes, we provide a concrete and orthogonal metric to reinforce that the reversed predictions in scVelo and standard VeloVI are not artifacts of visualization.

Furthermore, because count splitting yields two independent sets of spliced and unspliced count matrices, we can investigate the source of an RNA velocity method's low replicate coherence in more detail. We demonstrate this conflicting behavior of scVelo based on the two splits in the erythroid dataset (Fig. 2d). Note here we use the UMAPs computed from the split data rather than

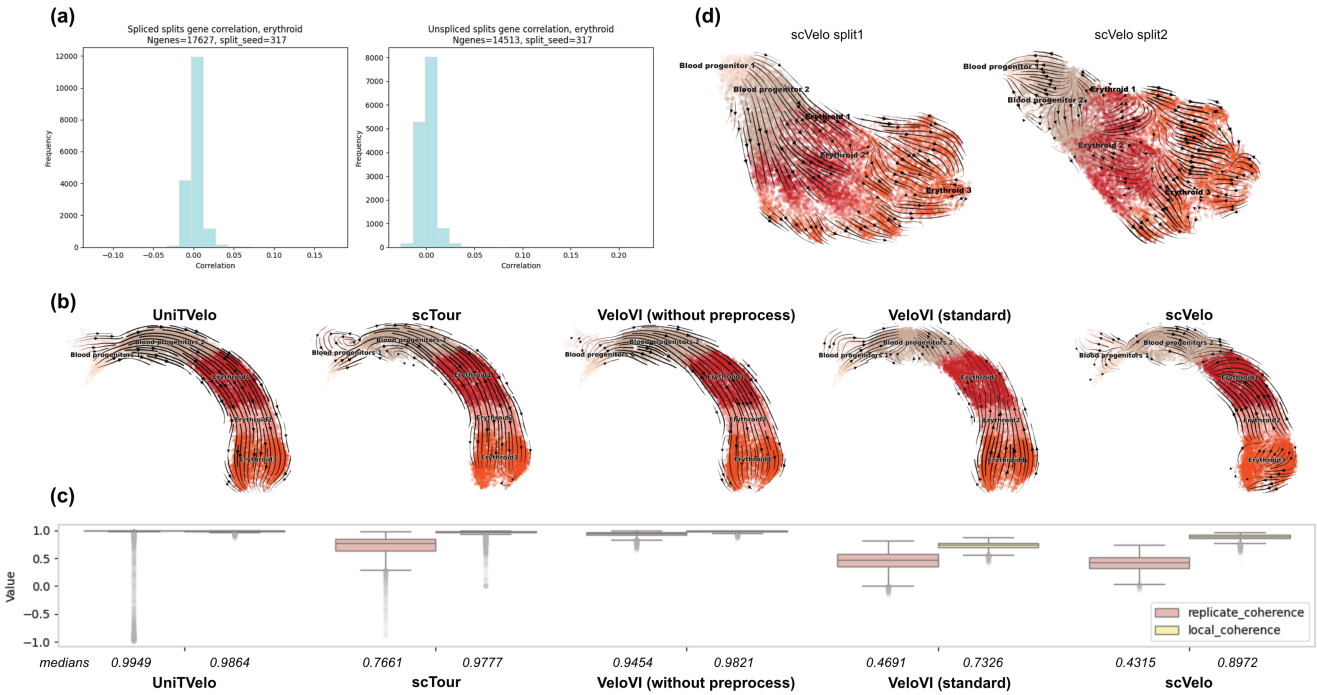

Fig. 2: Evaluation of replicate coherence on the erythroid dataset [Pijuan-Sala et al., 2019]. **(a)** The correlations of genes between two splits from the erythroid dataset are mostly around 0, indicating that count splitting effectively creates independent subsets that can reasonably be considered technical replicates. **(b)** The velocity estimates plotted on UMAPs for each of the five RNA velocity methods, with the embeddings directly inherited from the original dataset. **(c)** Boxplots showing local coherence of replicate coherence for each RNA velocity method results. **(d)** The velocity estimates of scVelo on two splits of the erythroid data, which show the velocity flow that conflicts with the true biology and has low replicate coherence values. The UMAP embeddings here are computed based on the split datasets with batch correction.

the UMAP attached in the raw data object as in Figure 2b. In Split 1, unlike the behavior on the entire dataset, the predictions appear surprisingly reasonable. However, Split 2 shows a more disorganized pattern and deviates from both Split 1 and the true trajectory, resulting in low replicate coherence. The full details of our analysis can be found in Supplementary Information S3, and additional plots of the RNA velocity flows using other methods are shown in Supplementary Figure S2.

3.2. Robustness of replicate coherence with unobserved time intervals and across intrinsic randomness

In reality, intermediate cell states along the developmental trajectory may sometimes be underrepresented due to factors such as their short lifespan, low abundance, or technological limitations. Prior work investigating RNA velocity [Atta et al., 2022, Li, 2023, Zheng et al., 2023] has raised concerns about this issue. Because the RNA velocity vectors are estimated from all the cells, it is possible that the estimated RNA velocity vectors are over-representing the transitions among cell states observed in the dataset. In this section, we evaluate the effectiveness of our framework in handling missing intermediate cell states by removing one intermediate cell state from a dataset and applying the workflow to both the original and modified datasets using five RNA velocity methods. We want to investigate if our replicate coherence is detrimentally affected if certain cell states are artificially removed from the analysis. In addition, due to the random nature of count splitting, concerns may arise about variability in results across different random seeds. We repeat the analysis across five seeds and show that the results are not dramatically affected.

We use the mouse pancreas dataset [Bastidas-Ponce et al., 2019] to demonstrate how a method's replicate coherence is preserved, even when cells at important developmental times are purposefully removed before the analysis. This dataset contains 3696 cells and 27,998 genes, where, critically, pre-endocrine cells are a necessary intermediary cell state that cells must go through before becoming alpha, beta, delta, and epsilon cells (Fig. 3a). Inspired by Li [2023], we ask how our replicate coherence behaves when intermediary cell states are omitted. We are interested in this because an ideal quantification of a cell's velocity uncertainty should remain robust even when an intermediary cell state is omitted. Therefore, our experiment first run all five methods on the original dataset to obtain velocity estimates. Next, we remove all pre-endocrine cells and re-analyzed each method on this modified dataset. We then compare the behavior of the methods and their replicate coherence under these two data settings. Visually, all five methods preserve similar predicted velocity flows on the UMAP embeddings in both datasets. For instance, Figure 3a displays RNA velocity estimates from scVelo on the two datasets. The velocity estimates from the other four methods can be found in Supplementary Figure

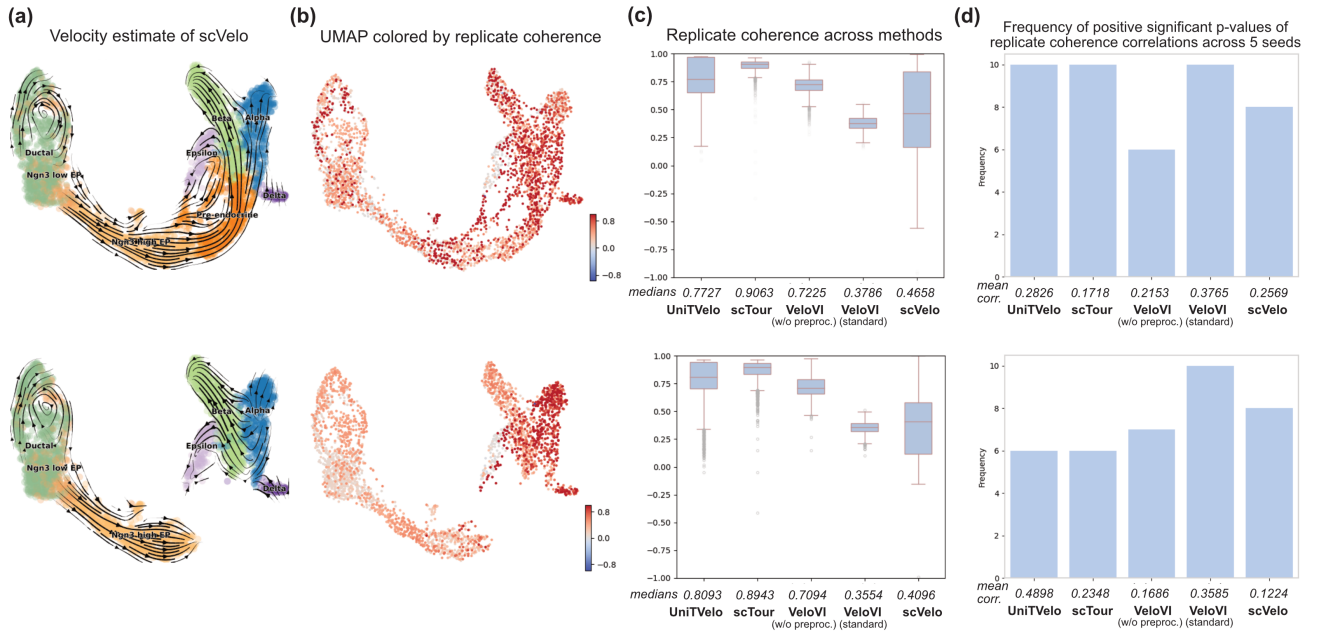

Fig. 3: Robustness of replicate coherence demonstrated on the pancreas data with and without pre-endocrine cells [Bastidas-Ponce et al., 2019]. **(a)** Velocity estimates of scVelo on UMAPs of pancreas data and pancreas data without pre-endocrine cells. **(b)** UMAPs of pancreas data and pancreas data without pre-endocrine cells colored by replicate coherence values. **(c)** Boxplots showing replicate coherence across five RNA velocity methods across five RNA velocity methods on the pancreas data with and without pre-endocrine cells, with the median replicate coherence values indicated at the bottom. **(d)** Frequencies of positive correlations with significant p-values for replicate coherence correlations computed across five seeds for five RNA velocity methods, with the mean correlation reported at the bottom.

S5. Figure 3b shows the replicate coherence for all methods on both datasets, with and without the pre-endocrine cells. We see that the replicate coherence is comparable in terms of the relative ordering of different methods across both datasets. Since the velocity estimates are similar between the pancreas dataset with and without pre-endocrine cells, it is reasonable to expect similar stability in the performance of each method. The replicate coherence results reflect this aspect, demonstrating its robustness even in the presence of unobserved cell states. For comparison, we also report VeloVI's intrinsic and extrinsic uncertainty quantification in Supplementary Figures S3 and S4. Additional plots visualizing the local and replicate coherence are in Supplementary Figures S6 and S7.

Next, we examine whether count splitting's intrinsic randomness in generating the count matrix splits affects the reliability of our replicate coherence. Similar to our previous analysis, an ideal quantification of a cell's velocity uncertainty should be robust to how the splits were exactly constructed. To investigate this, we split the pancreas dataset (with and without the pre-endocrine cells) five times, each setting the random seed to a different value. We then apply each RNA velocity method to the five pairs of count matrices. We compute replicate coherence on each pair of splits for each seed within a method and test the significance of Pearson's correlations for the replicate coherence values. In particular, since there are five different replicate coherence values for any particular RNA velocity method, we perform $\binom{5}{2} = 10$ hypothesis tests assessing if the correlation among the cell's replicate coherence between any two sets of results is significantly positive. Our results show that, for all methods, most correlations are positive and associated with significant p-values at a significance level of 0.05, using the Bonferroni correction (Fig. 3d). Furthermore, all methods exhibited mean correlations well above 0. This indicates that despite the randomness in this framework, replicate coherence remains robust across different seeds for each method. The replicate coherence correlations across multiple seeds are plotted in pairs in Supplementary Figures S9 through S13. We only performed five seeds due to limitations in computational resources – see Supplementary Table S3 for details on the time needed to run each method.

3.3. Extension to assess the biological pathways enriched by RNA velocity analyses

In this last section, we demonstrate that count splitting can aid in investigating RNA velocity methods beyond quantifying the velocity uncertainty. In particular, RNA velocity methods such as scVelo, UniTVelo, and standard veloVI have a two-step approach. These methods first fit an ODE RNA velocity model on each gene, and then they select a subset of genes with the best fit for all downstream velocity calculations. They adopt this two-step approach because not every gene is expected to exhibit the splicing dynamics, so only the relevant genes are kept. Given this understanding, we expect a good RNA velocity fit to select genes in biological pathways relevant to the underlying biology. We can use count splitting to investigate this hypothesis by continuing the analysis in Section 3.2 regarding the five seeds. Specifically,

for a given RNA velocity, we perform a gene set enrichment analysis (GSEA, Subramanian et al. [2005]) based on tabulating the number of each gene is selected across all ten instances the RNA velocity method was used (i.e., five seeds, and two splits per seed). We break ties based on the average model fit for that gene. We would expect that since scVelo, UniTVelo, and standard veloVI can capture the pancreas endocrinogenesis, each method would be enriched for pathways relevant to this biological system. We do not investigate this for scTour since it does not select genes, and veloVI without the preprocessing step uses the same genes as UniTVelo.

Our results show that scVelo, UniTVelo, and standard veloVI were all enriched for pathways related to pancreas endocrinogenesis. Specifically, after running a GSEA for each RNA velocity method, we select all the pathways with an adjusted p-value of less than 0.1. There are ten significant biological pathways across the three methods: GO:0051046 (regulation of secretion), GO:1903530 (regulation of secretion by cell), GO:0032940 (secretion by cell), GO:0023061 (signal release), GO:0046903 (secretion), GO:0002790 (peptide secretion), GO:0140352 (export from cell), GO:0030072 (peptide hormone secretion), GO:0015833 (peptide transport), and GO:0046879 (hormone secretion). These pathways collectively are about the differentiation and maturation of pancreatic endocrine cells, equipping them to produce and regulate the secretion of hormones essential for blood glucose control. They are fundamental to the pancreas's role in systemic metabolic regulation. Hence, using count splitting across multiple seeds, we can infer that each of the three RNA velocity methods identified appropriate genes to model this biological system. See Supplementary Information S3 for details about the method and Supplementary Figures S14 and S15 for additional results. Our analysis here showcases that the count splitting framework has broad applicability for biologists to assess how reliable an RNA velocity method is for their biological system.

4. Discussion

In this paper, we present a framework to assess the stability in the predicted velocity vectors of RNA velocity methods that applies to any RNA velocity methods. By creating two independent sets of hypothetical twin cells through count splitting, we can compare the method's results on two technical replicates of the same biological process. We applied our framework to five RNA velocity methods on two datasets, demonstrating its effectiveness in distinguishing between correct and incorrect method performance in mouse erythroid development and its robustness to unobserved time intervals using the pancreatic endocrinogenesis dataset.

Notably, our framework provides a generic way to measure the stability of any RNA velocity method. Unlike VeloVI's intrinsic and extrinsic uncertainty, which rely on posterior sampling and are specific to that method, our approach offers greater flexibility in assessing and comparing the performance of different methods. Additionally, since count splitting creates twin datasets via random sampling, it creates different splits and can lead to variations in replicate coherence across multiple runs. Compared to local coherence, our framework allows for the computation of a distribution of replicate coherence values. In our study, we run each method on each dataset five times to ensure the replicate coherence computations have similar behaviors. Ideally, this process can be repeated more times, given that the run time is reasonable.

While count splitting inevitably reduces the amount of sequence reads in the two generated splits compared to the original dataset, this mirrors real-world scenarios where low sequencing depths occur due to practical constraints. Nonetheless, it is reasonable to expect that a reliable RNA velocity method should produce similar results on the two splits, just as it would when applied to two datasets from the same biological process with low sequencing depths.

In our work, we compute replicate coherence on cell-specific velocity vectors based on two splits generated by count splitting. We focus on the high-dimensional velocity vectors since they are typically the most direct output from an RNA velocity method, unlike other derived quantities, such as the velocity graph and transition matrix, which involve additional smoothing. However, because some RNA velocity methods, such as scVelo and UniTVelo, internally select a subset of genes to model, the exact set of genes used for velocity computation could differ between two splits despite using the same RNA velocity method. Supplementary Table S2 summarizes the number of genes in the intersection and union of the gene sets in two splits for each method. To address this, we compute cosine similarity based on the union of genes from both splits. For genes present in one split but absent in the other, we pad the corresponding entries with 0. This approach ensures that all genes used for velocity computation in either split are considered while penalizing cases where genes are selected in only one split. Consequently, if a method selects very different sets of genes for the two splits, the replicate coherence will likely be lowered by the padded 0.

One potential limitation of our current work is that we applied the framework to two well-characterized scRNA-seq datasets. However, real-world datasets often present greater complexity. For instance, data may be sparse due to the nature of certain biological processes, or developmental trajectories may involve more complicated branching structures. How well our framework performs under such challenging conditions remains uncertain. Future work could focus on testing the framework with more diverse datasets that reflect real-world limitations, identifying the factors that most influence its performance, and refining the approach to handle these complexities effectively.

In addition, our current focus on boxplots and the median of replicate coherence values offers a more qualitative comparison rather than a formal statistical test. Such an approach lacks the ability to quantify the extent of differences in stability between methods or assess the significance of the difference. In Supplementary Table S4, we compare the replicate coherence, which is the paired cell-specific velocity cosine similarity, to a shuffled-pair version, which serves as a noninformative distribution. While we aim to account for the potential effect of velocity similarity across cells and capture the similarity of velocities between twin cells more precisely, we did not formalize this comparison as a test. A potential future direction is to incorporate replicate coherence into a hypothesis testing framework, enabling more quantitative comparisons and providing practical guidance for selecting RNA velocity methods.

Code and data availability

The code for replicate coherence (and negative-binomial count splitting) and all the analyses in this paper can be found at `https://github.com/linnykos/veloUncertainty`. The mouse erythroid dataset [Pijuan-Sala et al., 2019] is from `https://scvelo.`

readthedocs.io/en/stable/scvelo.datasets.gastrulation_erythroid.html, and the mouse pancreatic endocrinogenesis dataset [Bastidas-Ponce et al., 2019] is from https://scvelo.readthedocs.io/en/stable/scvelo.datasets.pancreas.html, both from the scVelo [Bergen et al., 2020] codebase.

Competing interests

No competing interest is declared.

Author contributions statement

YL, ZJW, YC, and KZL all contributed to the ideas behind this paper. YL and ZJW developed the code for the method. YL and KZL performed the analyses and wrote the paper.

Acknowledgments

We thank Lucy Gao, Dan Kessler, Anna Neufeld, and Daniela Witten for the wonderful discussions regarding count splitting that helped us solidify our ideas in this paper.

Supplemental data

The supplement provides more details of our analyses. Additionally, it includes extended visualizations and comparisons of our results, such as velocity estimates plotted on UMAPs colored by cell type, local coherence, and replicate coherence.

References

C. Ahlmann-Eltze and W. Huber. glmGamPoi: Fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics*, 36(24):5701–5702, 2020.

S. Ancheta, L. Dorman, G. Le Treut, A. Gurung, L. A. Royer, A. Granados, and M. Lange. Challenges and progress in RNA velocity: Comparative analysis across multiple biological contexts. *bioRxiv*, pages 2024–06, 2024.

L. Atta, A. Sahoo, and J. Fan. VeloViz: RNA velocity-informed embeddings for visualizing cellular trajectories. *Bioinformatics*, 38(2):391–396, 2022.

M. Barile, I. Imaz-Rosshandler, I. Inzani, S. Ghazanfar, J. Nichols, J. C. Marioni, C. Guibentif, and B. Göttgens. Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biology*, 22:1–22, 2021.

A. Bastidas-Ponce, S. Tritschler, L. Dony, K. Scheibner, M. Tarquis-Medina, C. Salinno, S. Schirge, I. Burtscher, A. Böttcher, F. J. Theis, H. Lickert, and M. Bakhti. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146 (12), 2019.

V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, 2020.

V. Bergen, R. A. Soldatov, P. V. Kharchenko, and F. J. Theis. RNA velocity: Current challenges and future perspectives. *Molecular Systems Biology*, 17(8):e10282, 2021.

A. Dharamshi, A. Neufeld, K. Motwani, L. L. Gao, D. Witten, and J. Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, pages 1–26, 2024.

M. Gao, C. Qiao, and Y. Huang. UniTVelo: Temporally unified RNA velocity reinforces single-cell trajectory inference. *Nature Communications*, 13(1):6586, 2022.

A. Gayoso, P. Weiler, M. Lotfollahi, D. Klein, J. Hong, A. Streets, F. J. Theis, and N. Yosef. Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature Methods*, 21(1):50–59, 2024.

G. Gorin, M. Fang, T. Chari, and L. Pachter. RNA velocity unraveled. *PLOS Computational Biology*, 18(9):e1010492, 2022.

Z. R. Hettinger, S. Hu, H. Mamiya, A. Sahu, H. Iijima, K. Wang, G. Gilmer, A. Miller, G. Nasello, A. D'Amore, D. A. Vorp, T. A. Rando, J. Xing, and F. Ambrosio. Dynamical modeling reveals rna decay mediates the effect of matrix stiffness on aged muscle stem cell fate. *bioRxiv*, pages 2023–02, 2023.

G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

J. Leiner, B. Duan, L. Wasserman, and A. Ramdas. Data fission: Splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

Q. Li. scTour: A deep learning architecture for robust inference and accurate prediction of cellular dynamics. *Genome Biology*, 24(1):149, 2023.

V. Marot-Lassauzaie, B. J. Bouman, F. D. Donaghy, Y. Demerdash, M. A. G. Essers, and L. Haghverdi. Towards reliable quantification of cell state velocities. *PLoS Computational Biology*, 18(9):e1010031, 2022.

A. Neufeld, J. Popp, L. L. Gao, A. Battle, and D. Witten. Negative binomial count splitting for single-cell RNA sequencing data. *arXiv preprint arXiv:2307.12985*, 2023.

A. Neufeld, L. L. Gao, J. Popp, A. Battle, and D. Witten. Inference after latent variable estimation for single-cell RNA sequencing data. *Biostatistics*, 25(1):270–287, 2024.

B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. Tyser, D. L. L. Ho, S. Srinvias, B. D. Simons, J. Nichols, J. C. Marioni, and B. Gottgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745): 490–495, 2019.

C. Qiu, J. Cao, B. K. Martin, T. Li, I. C. Welsh, S. Srivatsan, X. Huang, D. Calderon, W. S. Noble, C. M. Disteche, others, S. A. Murray, M. Spielmann, C. B. Moens, C. Trapnell, and J. Shendure. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nature Genetics*, 54(3):328–341, 2022.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

S. C. Zheng, G. Stein-O'Brien, L. Boukas, L. A. Goff, and K. D. Hansen. Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biology*, 24 (1):246, 2023.