Inferring time-lagged causality using the derivative of single-cell expression

Huan-Huan Wei[1], Hui Lu[1*], Hongyu Zhao[2*]

1 SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai, China
2 Department of Biostatistics, Yale University, New Heaven, USA

* huilu@sjtu.edu.cn, honyu.zhao@yale.edu

Abstract

Many computational methods have been developed for inferring causality among genes using cross-sectional gene expression data, such as single-cell RNA sequencing (scRNA-seq) data. However, due to the limitations of scRNA-seq technologies, time-lagged causal relationships may be missed by existing methods. In this work, we propose a method, called causal inference with time-lagged information (CITL), to infer time-lagged causal relationships from scRNA-seq data by assessing conditional independence between the changing and current expression levels of genes. CITL estimates the changing expression levels of genes by "RNA velocity". We demonstrate the accuracy and stability of CITL for inferring time-lagged causality on simulation data against other leading approaches. We have applied CITL to real scRNA data and inferred 878 pairs of time-lagged causal relationships, with many of these inferred results supported by the literature.

Author summary

Computational causal inference is a promising way to survey causal relationships between genes efficiently. Though many causal inference methods have been applied to gene expression data, none considers the time-lagged causal relationship, which means that some genes may take some time to affect their target genes with several reactions. If relationships between genes are time-lagged, the existing methods' assumptions will be violated. The relationships will be challenging to recognize. We demonstrate that this is indeed the case through simulation. Therefore, we develop a method for inferring time-lagged causal relationships of single-cell gene expression data. We assume that a time-lagged causal relationship should present a strong association between the cause and the effect's changing. To calculate such correlation, we first estimate the derivative of gene expression using the information from unspliced transcripts. Then, we use conditional independent tests to search gene pairs satisfying our assumption. Our results suggest that we could accurately infer time-lagged causal gene pairs validated by published literature. This method may complement gene regulatory analysis and provide candidate gene pairs for further controlled experiments.

Introduction

Single-cell RNA sequencing (scRNA-seq) is a technology capable of measuring the expression level of RNA at the single-cell resolution [1]. Rapidly growing scRNA-seq

data opens the door to a sufficiently powered inference of causality among genes. Several computational methods have been developed for causal inference from cross-sectional data (e.g., [2–4]) or time-series data (e.g., [5]). These methods have been applied with some success on biological data [6–8].

With reference to the time factor in causal inference, casual relationships among genes can be categorized into instant relationships and time-lagged relationships. In this study, we focus on the second. A time-lagged relationship is illustrated in Fig 1. The expression level of gene i at t_0 will affect the expression level of gene j at t_1, which is denoted by the black arrow connecting gene i with gene j. Note that with a time-lagged relationship, the expression level of gene i may not be related to the expression level of its target gene j at a specific time t_0.

Fig 1. Illustration of a time-lagged relationship across three time points. The gray rectangles represent different individual cells. Multi-trace measurements of three cells (top) and one cell's continuous measurements (bottom) are shown.

There are two main challenges to infer time-lagged causality on scRNA-seq data: the collection of longitudinal data and the presence of latent variables. First, it is difficult to continuously monitor the whole transcriptome within one cell. Of note, even when cells can be sequenced at different time points [9], such data cannot be considered as real time-series data because they capture different cells instead of the same set of cells. In Fig 1, the connections between time points are broken because distinct cell populations are studied. That is, we are not able to trace the evolutions of cells across different time points. We refer to such data as multi-trace data, where cells are collected from different time points. We will investigate whether such data may help us infer causality among genes through simulation studies. The reason why continuous measures matter is that there are natural confounders in inferring time-lagged causality on cross-sectional scRNA-seq data. For every cell, only the expression levels of genes (the colored ovals in the bottom part of Fig 1) at time point t_1 can be obtained from sequencing. For time-lagged relationships, the expression levels of the causal genes at the previous time point, i.e., t_0, act as confounders between the current expression levels of the causal genes and their targets' expression levels. As shown in Fig 1, the time-lagged causal gene pairs are not linked directly. If the expression levels of causal genes at previous time points are not considered, the association between the current expression levels of the causal genes and their targets can be low or even in the opposite direction. Throughout this paper, we refer to such confounders as natural confounders. This problem was noted previously [10] but has not been well addressed in the existing literature.

The second challenge is that unmeasured variables, also referred to as latent variables, are common in scRNA-seq experiments. scRNA-seq can capture the expression levels from 2000 to 6000 genes in a cell, where many genes with low-expression levels may not be captured. Besides, the causal path from one gene to another often involves many biological molecules which cannot be detected by scRNA-seq, such as proteins, metal ions, and saccharides. Together with low-expression genes, these latent variables are common for scRNA-seq data. However, many existing methods for causal inference assume the absence of latent variables, and as a result, may have difficulty in inferring causality from scRNA-seq data.

Here, we propose CITL (causal inference with time-lagged information), a method to infer the time-lagged causal relationships among genes in scRNA-seq data capable of overcoming the two challenges mentioned above. CITL uses RNA velocity information inferred from scRNA-seq data to estimate the changing expression levels of genes. By assessing conditional independence between the changing and current expression levels, CITL can more accurately infer time-lagged relationships than a commonly-used

cross-sectional causal inference algorithm, the PC-stable algorithm [4] in simulations. Compared with [8], which also uses RNA velocity to infer causality, CITL is more stable in simulation studies and may better identify time-lagged causality from extensive real data. On real scRNA-seq data, we show the concordance between the time-lagged causal relationships inferred by CITL and regulatory pathways curated by published literature. Our results also suggest that time-lagged causality may represent the relationships involving multi-modal variables.

Materials and methods

Causal inference with time-lagged information (CITL)

We make the following assumption for our causal inference:

Time-Lagged Assumption: if the current expression level of gene i X_i^{cur} is strongly correlated with the changing expression level of gene j X_j^{cha}, then gene i is inferred to be the cause of gene j in a time-lagged manner.

A strong correlation means that X_i^{cur} and X_j^{cur} are dependent conditioning on other variables, which can be assessed by the conditional independence (CI test). With this assumption, the X^{cha} of a gene is not related to its X^{cur} value but is correlated with the X^{cur} of its causal genes. Therefore, the X_i^{pre}, the natural confounder between X_i^{cur} and X_j^{cur}, does not directly influence X_j^{cha}. X_i^{pre} can influence X_j^{cha} only through X_i^{cur}, which means it is not a natural confounder for the correlation between X_i^{cur} and X_j^{cha}.

RNA velocity [11] offers a way to estimate gene expression changes based on spliced mRNA and unspliced RNA information. CITL uses RNA velocity for a unit of time as the changing expression level X^{cha} and the extrapolated expression levels in a unit of time as the subsequent expression level X^{sub}. Note that we use a fixed unit time in this manuscript as an approximation, although the length of time that different genes exert effects on other genes may differ. For consistency, we used the same parameters described in [11] to calculate RNA velocity.

To infer time-lagged causal relationships, CITL first constructs an undirected graph (UG) through both X^{cur} and X^{cha}. Each node in the UG represents the X^{cur} or X^{cha} of a gene. Each edge in the UG represents the dependency between the X^{cur} (or X^{cha}) of a gene and that of another gene. The dependency is assessed by CI test conditional on at most k ($\leq$ the number of nodes n) genes. CITL then focuses on the edges linking the X^{cur} of some genes to the X^{cha} of others. If the X^{cur} of a gene is linked to the X^{cha} of another in the UG, the former gene is assigned as the cause of the latter gene. We note that the X^{cur} (or X^{cha}) of some genes can be linked to each other. We assume that these connections do not represent time-lagged relationships. Thus they are not the focus of this work. We provide an open-source command-line tool of CITL at https://github.com/wJDKnight/CITL.

Comparisons with other methods

We compared the performance of CITL versus a commonly-used causal Bayesian network method, PC-stable [4], and a recently published causal inference method for scRNA-seq data [8], Scribe. PC-stable first constructs a UG as well. Therefore, CITL adopts the same strategy to construct the UG by using the *bnlearn* package [12]. The difference is that PC-stable uses probabilistic dependency to determine causal direction under three assumptions: *Causal Sufficiency*, *Causal Markov Assumption*, and *Faithfulness* [2,13]. We compare the performance of CITL with the PC-stable through simulations under different approaches of analyzing scRNA-seq data as detailed in the following.

- Approach 1: PC-stable, using only X^{cur}. This is the simple adoption of the causal inference methods to scRNA-seq data. As discussed above, it will not be able to infer time-lagged relationships. We include this approach to assess the lack of power to identify time-lagged relationships with only X^{cur}.

- Approach 2: PC-stable, using X^{cur} and X^{sub}, where X^{sub} is the extrapolated expression levels at the subsequent time point. For this approach, PC-stable is applied to time-lagged data. However, natural confounders may still exist between X^{cur} and X^{sub}. Consequently, we consider this scenario to assess the effect of natural confounders on causal inference.

- Approach 3: PC-stable, using X^{cur} and X^{cha} but without *time-lagged assumption*. This approach infers causality by PC-stable itself based on PC-stable's assumptions. We include this approach to investigate the usefulness of *time-Lagged Assumption*.

We note that any method which can identify a strong correlation between X^{cur} and X^{cha} may be suitable for the proposed framework. In addition to the above three approaches, we also consider another approach, Approach 0, which is the simplest version of the proposed framework using Pearson's correlation coefficient to discover a strong correlation between X^{cur} and X^{cha}. If the absolute value of Pearson's correlation coefficient between X_i^{cur} and X_j^{cha} is above a threshold, we infer gene i as the cause of gene j as baseline prediction.

We also consider a recently published causal inference method for scRNA-seq data [8], Scribe. It uses restricted directed information (RDI) to evaluate the causal effect of the current expression levels on the subsequent expression levels. Similar to Approach 2, Scribe assumes that if the RDI of X_i^{cur} and X_j^{sub} is higher than a threshold, gene i is the cause of j. The default values of the parameters of Scribe were used in simulation studies.

Simulation

Some experiments sequence cells at one time point while others sequence cells at multiple time points. We refer to the former as single-trace data and the latter as multi-trace data. We considered both scenarios in our simulations. For single-trace data, we simulated 3000 cells. For multi-trace data, we simulated from three traces with each trace having 1000 cells. We carried out 500 simulations for each set-up. For each simulation, we randomly generated a causal graph G_{true} that contained 50 nodes (genes) and 50 directed edges on average. The probability of an edge between nodes was 4.1%, and its direction was randomly assigned. Time-lagged relationships were simulated in the following manner:

$$\begin{aligned} X_i^{cur} &= f_1(X_i^{pre}) + f_2(causalpre(X_i)) + e^{cur} \\ X_i^{sub} &= f_1(X_i^{cur}) + f_2(causalcur(X_i)) + e^{sub} \\ X_i^{cha} &= X_i^{sub} - X_i^{cur} \end{aligned} \tag{1}$$

For each cell, we simulated four values related to each of the 50 genes' expression levels, including previous X_i^{pre}, current X_i^{cur}, subsequent X_i^{sub}, and changing X_i^{cha}. Based on the collected values of X_i^{pre} and the causal graph, X_i^{cur}, X_i^{sub}, and X_i^{cha} were generated through Eq (1) using $causalpre(X_i)$ (the previous values of the causes of X_i) and $causalcur(X_i)$ (the current values of the causes of X_i). e^{cur} and e^{sub} represent standard Gaussian noise $N(0, 1)$.

Here, we used linear functions to describe time-lagged relationships. The coefficient of X^{pre} or X^{cur} in the linear function $f_1()$ was 0.8, simulating the transcripts of genes that spontaneously degrade over time. The coefficients of all causal genes in $f_2()$ were set to 1, assuming all causal genes had the same effect on their effector genes. In addition, we assumed that X_i^{pre} did not interact with $causalpre(X_i)$, meaning there was no feedback, and we could add $f_1(X_i^{pre})$ and $f_2(causalpre(X_i))$.

For single-trace data, X_i^{pre} was assumed to follow a log-normal distribution with a constant mean and standard deviation ($\ln(X) \sim N(0, 0.04)$). We chose a log-normal distribution because RNA sequencing data are often skewed, rather than like a normal distribution. For every run of multi-trace simulation, we simulated three separate sets of data (trace). In each trace, X_i^{pre} followed a log-normal distribution ($\ln(X) \sim N(\mu, 0.04)$) with its mean (μ) randomly drawn from a uniform distribution between 0 and 2. Then, the three traces were merged into one data set, taking into no account of the trace information. For simulations considering latent variables, we randomly removed a certain proportion of the nodes (genes) after data generation.

In our simulations, we also investigated whether CITL can infer non-time-lagged relationships, referred to as instant causal relationships. This assumes that the current expression level of a gene results from its previous expression level and the current expression level of its causes. These data were generated in a similar manner as the time-lagged except for the method used to generate X_i^{cur} and X_i^{sub}. For instant simulation, we considered Eq 2(2), where $causalsub(X_i)$ is the subsequent values of the causes of X_i.

$$\begin{aligned} X_i^{cur} &= f_1(X_i^{pre}) + f_2(causalcur(X_i)) + e^{cur} \\ X_i^{sub} &= f_1(X_i^{cur}) + f_2(causalsub(X_i)) + e^{sub} \end{aligned} \tag{2}$$

To equally benchmark CITL against Scribe, the simulation data in [8] were used. The simulation was based on a core network of neurogenesis with 12 genes forming 13 directed pairs and two bidirectional pairs. Data were simulated according to the differential equations of these genes. We tested the performance of CITL and Scribe in simulations under both Qiu's and our set-ups. In all simulations, the k of the CI test was set to be equal to the square root of the number of genes n.

Evaluation

We used precision, recall, and F-measure for the inferred node adjacency versus the data generating model as the primary evaluation measures to compare the performance of different approaches. In addition, we used the ability of determining directions (ADD) to evaluate how well a method was able to define directions given true causal edges. To compute these metrics, we first calculated three basic statistics: true positives (TP), false positives (FP), and false negatives (FN) that are related to inferring edges. TP is the number of adjacencies in both the output graph G_{output} from an analytical approach and the true graph G_{true}. FP represents the number of adjacencies in G_{output} but not in G_{true}. FN is the number of adjacencies in G_{true} but not in G_{output}. Precision is the ratio TP/(TP+FP), recall is the ratio TP/(TP + FN), and F-measure is the ratio 2 * precision * recall / (precision + recall). For evaluating the directions, $\text{TP}_{direction}$ represented the number of directed edges in both G_{output} and G_{true} with consistent directions. FP represents the number of inconsistent edges in G_{output} compared with G_{true}, including absent, undirected, and reverse. FN represents the number of edges in G_{true} but not correctly directed in G_{output}. ADD was calculated by $\text{TP}_{direction}/\text{TP}_{edge}$.

Data sets

We considered two data sets. Data set 1 was from mouse P0 and P5 dentate gyrus [14], and RNA velocity information was estimated with the same parameters as the example dentate gyrus in Velocyto (http://velocyto.org/). There were more than 18,000 cells and an average of 2,160 genes for each cell in data set 1 after preprocessing. Data set 2 was the human week ten fetal forebrain data set in Velocyto, containing 1,720 cells and an average of 1,488 genes for each cell. According to La Manno *et al.* (2018), the forebrain, as identified by pre-defined markers, can be divided into eight developing stages (0-7). The stage information was only exploited in data visualization.

Results

Simulation results

Simulation results of Approach 0

The performance of Approach 0 largely depends on the threshold of Pearson's correlation coefficient. We tested its performance at 18 thresholds from 0.1 to 0.9 through 500 simulations for each setting. Fig 2 summarizes the performance of Approach 0 in single-trace simulations (top row in Fig 2) and multi-trace simulations (bottom row in Fig 2). In single-trace simulations, the precision increased, and the recall decreased, as the threshold increased for both finding edges and determining causal directions. The more stringent the threshold was, the more accurate Approach 0 was, but the fewer edges Approach 0 could find. When the threshold was around 0.2, Approach 0 achieved the highest F-measure in single-trace simulations. In contrast, the highest F-measure of Approach 0 in multi-trace simulations was achieved when the threshold was around 0.75. The overall performance of Approach 0 in multi-trace simulations was much worse than that in single-trace simulations. It suggests that multiple traces induce many false positives for both finding edges and determining causal directions in Approach 0.

Fig 2. Results of Approach 0 for single-trace simulations (top row) and multi-trace simulations (bottom row) at different thresholds of "the strong correlation".

Comparisons CITL with PC-stable and its variant approaches

The simulation results for the single-trace scenario for the other approaches are summarized in Table 1. For finding edges, Approach 0 achieved the lowest precision, which is expected as PC-stable applied to current expression levels will miss time-lagged causal edges via single-trace data. The recall of Approach 2 was lower than others, which suggests that the natural confounders in Approach 2 clearly influenced the discovery of casual edges. Approach 3 and CITL derived the same UG, which performed the best in both recall and F-measure, demonstrating that changing information is useful when identifying edges between causal pairs from single-trace data. When determining the causal direction, CITL performed best, and Approach 1 had the worst performance. Both Approach 1 and Approach 2 performed worse than Approach 3 in recall and F measure, indicating that natural confounders influence the determination of causal directions. CITL was better than Approach 3 for all three metrics, demonstrating that CITL was most effective in determining causal directions than the assumptions of PC-stable. As for multi-trace simulations, we obtained similar results, as

shown in Table A in S1 Appendix. Comparing the results of CITL with those of Approach 0, CITL outperformed Approach 0 in multi-trace simulations (Fig 2, Table A in S1 Appendix). In summary, CITL had the best performance among the approaches and was less sensitive to the type of data applied.

As for the results for inferring instant causality, the type of simulation affected the determination of directions. Comparing with results in single-trace simulations (Table B in S1 Appendix), the F-measure for determining the directions of Approach 1 decreased while that of CITL increased in multi-trace simulations (Table C in S1 Appendix). This suggests that natural confounders, the previous expression levels of causes, could have a larger effect on instant causal relationships for multi-trace data. In this case, CITL was still a good choice for identifying instant causality.

Table 1. Comparisons of different approaches based on PC-stable.

Edges	**Approach 1**	**Approach 2**	**Approach 3/CITL**	
Precision	0.651 (0.279)	0.975 (0.049)	0.974 (0.033)	
Recall	0.434 (0.275)	0.262 (0.081)	0.597 (0.104)	
F-measure	0.540 (0.264)	0.407 (0.099)	0.734 (0.081)	
Directions	**Approach 1**	**Approach 2**	**Approach 3**	**CITL**
Precision	0.201 (0.154)	0.838 (0.180)	0.573 (0.247)	0.859 (0.143)
Recall	0.137 (0.132)	0.224 (0.084)	0.334 (0.137)	0.512 (0.104)
F-measure	0.213 (0.124)	0.348 (0.113)	0.420 (0.169)	0.636 (0.108)

The average values from 500 single-trace simulations are shown with standard deviation values in parentheses.

Comparisons with Scribe under different simulation settings

We also evaluated the performance of Scribe. Since the runtime of Scribe for one simulation (about 20 minutes) was longer than others (a few seconds), we only applied Scribe to one replicate of each simulation. In single-trace simulation, Scribe calculated the RDI of the 2450 edges (all possible combinations of 50 nodes) and removed the edges with smaller RDI leaving 1225 edges where all 50 real edges and 25 real directions were captured. Among the top-100-RDI edges, about four edges were true causal relationships, suggesting that the performance of Scribe was not better than random. Similar results were obtained for multi-trace and instant simulations where Scribe could not reveal causal relationships from the simulated data. We further compared Scribe with CITL through simulations conducted as previously described by [8]. In the simulation, the standard deviation of the intrinsic noise in the differential equations was set to be equal to 0.01 or 2; this represented the randomness of the causal effect, the temporal fluctuation, and random error. The results are shown in Table 2. Under the low-noise setting, the top 9 RDI edges inferred by Scribe were better than the CITL results. On the other hand, CITL performed better under the high-noise setting; CITL discovered more true positive edges and directions than the top 19 of Scribe. The performance of Scribe under our simulation set-up and Qiu's set-up was very different. In contrast, CITL performed well in both sets of simulations.

Simulation with latent variables

To evaluate the impact of latent variables on CITL to infer time-lagged causality, we performed single-trace simulations by randomly removing 0%, 10%, 30%, and 50% of the total genes. As illustrated in Fig 3(a, b), as the number of latent variables increased, the performance of all approaches reduced for both finding causal edges and

Table 2. Comparisons between CITL and Scribe under the simulation setting of Qiu *et al.* with two different noise levels.

sd = 0.01	**CITL**	**Scribe(top 9)**	**Scribe(top 19)**
All edges	9	9	19
TP_{edges}	4	4	7
$TP_{direction}$	1	4	7
sd = 2	**CITL**	**Scribe(top 9)**	**Scribe(top 19)**
All edges	19	9	19
TP_{edges}	13	6	11
$TP_{direction}$	10	6	9

sd: standard deviation

determining causal directions. This showed that latent variables had a negative effect on all approaches as expected. CITL performed the best across all the simulation settings. We used ADD to evaluate how well an approach inferred the causal directions in the presence of latent variables. The distribution of ADD in the simulations is shown in Fig 3c. The ADD of CITL concentrated at a higher level, while other approaches were not stable. This shows that CITL is more robust than other approaches. Similar results were obtained for multi-trace simulations (Fig A in S1 Appendix).

Fig 3. Results of single-trace simulation with latent variables. a: The performance of discovering edges. b: The performance of determining directions. c: Ability of determining directions.

Applications to real data sets

Evaluation of the information in RNA velocity for inferring causal relationships

For real data sets, we estimated the changing expression levels and the subsequent expression levels by RNA velocity. Before adopting the estimated X^{cha} to infer causality, we investigated how much information it contained. First, we observed that using the estimated X^{cha} to calculate the correlation led to different correlated pairs than when using X^{cur} in data set 1 (Fig 4a) and data set 2 (Fig 4b). This suggests that the information for the estimated X^{cha} was different from that of X^{cur}. A similar method recently developed drew the same conclusion [15]. Second, we applied Approach 0 to both data sets, and the resulting networks showed that the distribution of indegree and outdegree was very different (Fig A in S2 Text). In addition, the molecular function of low-outdegree genes was associated with gene regulation (S2 Text). Taken together, the unique information of the estimated X^{cha} suggests that CITL could use RNA velocity to estimate the changing expression levels.

Causal inference using CITL on real data sets

We applied CITL to data set 1 and data set 2 with 2,508 and 878 time-lagged causal pairs (TLPs) inferred, respectively. We also applied PC-stable on the data sets with current-only expression data and compared the gene pairs inferred by PC-stable to TLPs. For computational efficiency, the value of k for both CITL and PC-stable was set to be equal to the square root of the number of genes for each data set. A total of 3,998 and 4,459 pairs were inferred by PC-stable from data set 1 and data set 2, respectively.

Fig 4. The distribution of intersections as measured by different methods in two data sets. We compared the most correlated genes derived from X^{cur}, X^{cha}, and X^{sub}. Pearson's correlation coefficient is used to describe three different correlations: [1] correlation between X_i^{cur} and X_j^{cur} (denoted as cur_cur), [2] correlation betwee X_i^{cur} and X_j^{sub} (denoted as cur_sub); and [3] correlation between X_i^{cur} and X_j^{cha} (denoted as cur_cha). For each gene, 100 genes with the largest (50 positive and 50 negative) cur_cur, cur_sub, or cur_cha values were collected into three gene sets S^{cc}, S^{cs}, and S^{ch}, respectively. We recorded the number of genes in the intersections between the three sets. In addition, we recorded the intersection between S^{cc} and 100 randomly selected genes as controls. a: The distribution in data set 1. b: The distribution in data set 2.

In data set 1, only four gene pairs were found by both approaches, and there was no overlap for data set 2. These results suggest that CITL infers different types of causality from previous methods that only used the current expression level of genes.

CITL accurately infer time-lagged causal pairs

Because we do not know the ground truth for time-lagged causality, we investigated the biological relationships of TLPs to evaluate the performance of different methods. Pathway Studio (http://www.pathwaystudio.com/) enables searching interactions between molecules, cell processes, and diseases from the literature. Almost any pair of two genes could be related, directly or indirectly, through Pathway Studio. Each interaction is annotated by a sentence from the literature. Not all interactions are regulatory, such as binding. We reviewed the annotation of every searched interaction to find TLPs with regulatory interactions. For the regulatory interactions, we divided them into two categories. The "PROT" type refers to interactions that only involve proteins, such as increasing or reducing protein activity, co-activating or antagonizing, and phosphorylating or dephosphorylating. The "TRSC" type refers to interactions relating to proteins regulating the transcription of specific genes, including activation and repression. Considering manually filtering interactions taking considerable time, we only investigated the biological functions of a subset of the pairs.

In the following, we describe how we chose the subset of TLPs to consider. Single-trajectory developmental cells in data set 2 are easier to visualize time-lagged relationships than multi-trajectory differentiating cells in data set 1. Therefore, we focus on the TLPs in data set 2, where 37 transcription factors were involved in 68 TLPs. Transcription factors (TFs) were taken from the TRRUST database, a repository of curated TF-target relationships of human and mouse [16]. We investigated these 68 TLPs in Pathway Studio and manually checked the interactions of each TLP.

All the 68 pairs had indirect relationships, forming paths with one or more intermediates. Most of the interactions among these paths were "non-regulatory". We focused on the regulatory paths ended with a TRSC interaction, since the causality among genes' transcripts, rather than proteins, was of interest in scRNA-seq. 14 TLPs with regulatory relationships (rTLPs) and their regulatory paths are shown in Table 3. The interaction types are listed from the left of the corresponding path to the right. CITL achieved an accuracy of 0.93 (13/14) for correctly inferring the causal directions of rTLPs. The regulatory effect (activation or repression) of 11 pairs were correctly described. Only one rTLPs was assigned an inconsistent direction with its path (the cur_cha of *MAGED1* – *EOMES* was -0.19).

To evaluate the significance of the accuracy of CITL, we first investigate how likely a random gene could be the target of a TF. We randomly chose 11 TFs from the 37 TFs and investigated their regulatory relationships with randomly selected genes. For each TF, a randomly selected gene was assigned as its effect. Then, the functional connection

Table 3. Detailed paths of the causal pairs with regulatory relationships.

rTLP	Path	cur_cur	cur_cha	Type of each step
TCF7L2 → *FTH1*	*TCF7L2* → *RELA* → *FTH1*	-0.98	0.96	PROT, TRSC
YBX1 → *EPS8*	*YBX1* → NF-κ-B signaling → *EPS8*	0.04	0.95	PROT, TRSC
NFIA ↛ *EIF4A2*	*NFIA* ↛ ERK pathways ↛ *PPARGC1A* ↛ *EIF4A2*	0.81	-0.98	PROT, TRSC, TRSC
SATB2 ↛ *UBB2B*	*SATB2* ↛ MAPK7 → NEUROG2 → *UBB2B*	0.71	-0.95	PROT, PROT, TRSC
VHL ↛ *ABCD2*	*VHL* ↛ WNT → AMPK → *ABCD2*	0.63	-0.58	PROT, PROT or TRSC, TRSC
VHL ↛ *PCDH9*	*VHL* → *TP53* ↛ *PCDH9*	0.39	-0.88	TRSC, TRSC
TSC22D1 ↛ *PTPRD*	*TSC22D1* ↛ MTOR ↛ MYCN ↛ *PTPRD*	0.03	-0.92	PROT, PROT, TRSC
TSC22D1 ↛ *ZBTB18*	*TSC22D1* → TGFB1 ↛ ASCL1 → *ZBTB18*	0.01	-0.84	PROT or TRSC, PROT, TRSC
MLLT3 ↛ *FLRT3*	*MLLT3* → E2F1 ↛ WNT → *FLRT3*	-0.19	-0.99	PROT, PROT, TRSC
NFKB1 → *HSPA8*	*NFKB1* → *MYB* → *HSPA8*	0.69	0.99	TRSC, TRSC
SFPQ ↛ *JUND*	*SFPQ* ↛ PGR → *JUND*	-0.67	-0.96	PROT, TRSC
HDAC2 ↛ *HSP90AA1*	*HDAC2* ↛ *MIR15A* ↛ *HSP90AA1*	-0.80	-1.00	TRSC, TRSC
TSC22D1 ↛ *KDM5B*	*TSC22D1* → transforming growth factor → *KDM5B*	0.40	-0.82	PROT, TRSC
EOMES → *MAGED1*	*EOMES* ← ERK1/2 ← *MAGED1*	-0.03	0.97	TRSC, PROT

cur_cur: the Pearson's correlation coefficient between the current expression levels of the cause and the target; cur_cha: the Pearson's correlation coefficient between the current expression level of the cause and the changing expression level of the target. →: up-regulation. ↛: down-regulation.

Fig 5. *MLLT3* → *FLRT3*. a: Scatter plot of the current expression levels of *MLLT3* and *FLRT3*. b: Scatter plot of the current expression level of *MLLT3* and the changing expression level of *FLRT3*. n: Box plots of the normalized current expression levels of *MLLT3* and *FLRT3* at eight stages, which was identified by pre-defined markers [11].

between the gene pair, referred to as randomly-selected-and-direction-assigned pair (RAP), was searched using Pathway Studio. Like the TLPs inferred by CITL, most RAPs did not have regulatory relationship. To find a gene having a regulatory relationship with each TF, we searched 35 RAPs. In the 11 RAPs with regulatory relationships (rRAPs), only two rRAPs' assigned directions were consistent with their known causal directions. Therefore, we speculate that, for a TF, there are more upstream genes than downstream after excluding non-regulatory genes. We compared the accuracy of CITL to the accuracy of random selection using Fisher's exact test. The p-value of the test was 0.00024, suggesting the excellent performance of CITL.

A example of time-lagged causal pair

We highlight a time-lagged causal pair, "*MLLT3* → *FLRT3*" in Fig 5. "*MLLT3* → *FLRT3*" is a gene pair with a small negative cur_cur correlation (−0.19) and a large negative cur_cha correlation (−0.99). Though the correlation between the current expression levels was weak, this gene pair showed a strong negative correlation in terms of time-lagged association. The inconsistency can be explained as follows. The decrease of *FLRT3* in stages 5 and 6 is due to the high expression level of *MLLT3* in stages 3 and 4 (Fig 5c). We further investigated whether this pair had a transcriptional causal relationship. *MLLT3* participated in the activity of E2F1 protein [17], which could repress WNT signaling [18]. WNT signaling could control the expression of *FLRT3* [19]. In short, *MLLT3* could repress the expression of *FLRT3*, which is consistent with the result of CITL.

CITL overcomes the limitations of scRNA-seq

Indirect regulations involved more biological reactions than direct regulations, making it more reasonable to consider time-lagged relationships. Due to technical limitations, some intermediates in the indirect regulations were difficult to be detected by scRNA-seq. Therefore, researchers often have to deal with indirect relationships. Here, the only path in which all genes were detected was "*YBX1* $\rightarrow$ NF-κ-B $\rightarrow$ *EPS8*". The protein encoded by *YBX1* can activate NF-κ-B signaling [20], which then induces the transcription of EPS8 [21]. The cur_cur correlations between "*YBX1* – *NFKB1*", "*NFKB1* – *EPS8*", and "*YBX1* – *EPS8*" were -0.72, -0.70, and 0.04, respectively. None of these could explain the relationship between *YBX1* and *EPS8* in the literature. On the other hand, the cur_cha correlation between "*YBX1* – *EPS8*" was 0.95, consistent with the relationship between the genes. The results demonstrates that indirect relationships can be time-lagged relationships and that CITL is a better way of discovering these relationships.

Furthermore, some intermediates were not RNA at all. As shown in Table 3, most paths involved PROT steps. The best way to describe "*YBX1* $\rightarrow$ *EPS8*" would need the expression level of *YBX1*, the protein activity of NF-κ-B and the expression level of *EPS8*. Although many single-cell multi-omics technologies have been developed, none of these can ensure that all of the necessary molecules in each cell are quantified. However, CITL accurately inferred indirect relationships without any protein-level information. Consequently, the CITL, discovering time-lagged relationships, was more practical than previous methods which focused on instant interactions in scRNA-seq data.

Discussion

The changing expression levels of genes are crucial to CITL. Thus, the approach used to estimate these levels can have major impact on the results. There are two main challenges to correctly estimate the changing expression levels with RNA velocity. First, scRNA-seq technologies have limitations on quantifying transcripts. The quality of raw data is of great importance to results. Second, the inference of RNA velocity depends on some tuning parameters [11]. There is no gold standard to evaluate the estimated changing expression levels. Despite the two obstacles, RNA velocity has proved its usefulness to estimate transcriptional changes of genes in many applications [22, 23]. Also, Qiu *et al.* investigated three approaches to deriving single-cell time-series data and concluded that RNA velocity was the most appropriate way to estimate real time-series data through simulations [8].

A drawback of CITL is that it cannot distinguish whether the type of relationships is time-lagged or instant. In biology, the relationships between genes can be a mixture of time-lagged and instant relationships. If we can confirm the interactional type of each gene pair and adapt CITL to the type, the overall accuracy may be greatly improved.

Conclusion

In this article, we propose CITL to infer the time-lagged causality of genes using scRNA-seq data. Specifically, we adopt the changing information of genes estimated by RNA velocity in our approach. We further present the superior performance of CITL against other methods in simulations under different set-ups. The proposed approach CITL achieves promising results on a human fetal forebrain scRNA-seq data set, which accurately provides time-lagged causal gene pairs curated by published articles. We note that most methods for analyzing scRNA-seq data did not consider the relationships between genes that could be time-lagged. The results of simulations and

real data sets from this paper suggest that we cannot ignore such common relationships. Therefore, we foresee that CITL can provide more insights that may help to guide future gene regulatory research.

Supporting information

S1 Appendix. Supplementary simulation results of different set-ups.

S2 Text. The degree distribution of genes and their molecular function.

Acknowledgments

This work was supported by the National Key R&D Program of China [2018YFC0910500], the Neil Shen's SJTU Medical Research Fund, and SJTU-Yale Collaborative Research Seed Fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82. doi:10.1038/nmeth.1315.
2. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search, 2nd Edition. MIT press; 2001.
3. Ramsey J, Zhang J, Spirtes PL. Adjacency-Faithfulness and Conservative Causal Inference. arXiv e-print. 2012; p. 1206.6843.
4. Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. Journal of Machine Learning Research. 2014;15(1):3741–3782.
5. Runge J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. 2018;28(7):075310. doi:10.1063/1.5025050.
6. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523–9. doi:10.1126/science.1105809.
7. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 2013;153(3):707–720. doi:10.1016/j.cell.2013.03.030.
8. Qiu X, Rahimzamani A, Wang L, Ren B, Mao Q, Durham T, et al. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. Cell Syst. 2020;10(3):265–274.e11. doi:10.1016/j.cels.2020.02.003.
9. Hrvatin S, Hochbaum DR, Nagy MA, Cicconet M, Robertson K, Cheadle L, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. Nat Neurosci. 2018;21(1):120–129. doi:10.1038/s41593-017-0029-5.
10. Epskamp S, Waldorp LJ, Mõttus R, Borsboom D. The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. Multivariate Behav Res. 2018;53(4):453–480. doi:10.1080/00273171.2018.1454823.

11. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494–498. doi:10.1038/s41586-018-0414-6.

12. Nagarajan R, Scutari M, Lèbre S. Bayesian Networks in R. vol. 48. Springer; 2013.

13. Pearl J. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge University Press; 2009.

14. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular Architecture of the Mouse Nervous System. Cell. 2018;174(4):999–1014.e22. doi:10.1016/j.cell.2018.06.021.

15. Eghbalnia HR, Wilfinger WW, Mackey K, Chomczynski P. Coordinated analysis of exon and intron data reveals novel differential gene expression changes. Sci Rep. 2020;10(1):15669. doi:10.1038/s41598-020-72482-w.

16. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018;46(D1):D380–d386. doi:10.1093/nar/gkx1013.

17. Hallstrom TC, Mori S, Nevins JR. An E2F1-dependent gene expression program that determines the balance between proliferation and cell death. Cancer Cell. 2008;13(1):11–22. doi:10.1016/j.ccr.2007.11.031.

18. Hughes TA, Brady HJ. Cross-talk between pRb/E2F and Wnt/beta-catenin pathways: E2F1 induces axin2 leading to repression of Wnt signalling and to increased cell death. Exp Cell Res. 2005;303(1):32–46. doi:10.1016/j.yexcr.2004.09.014.

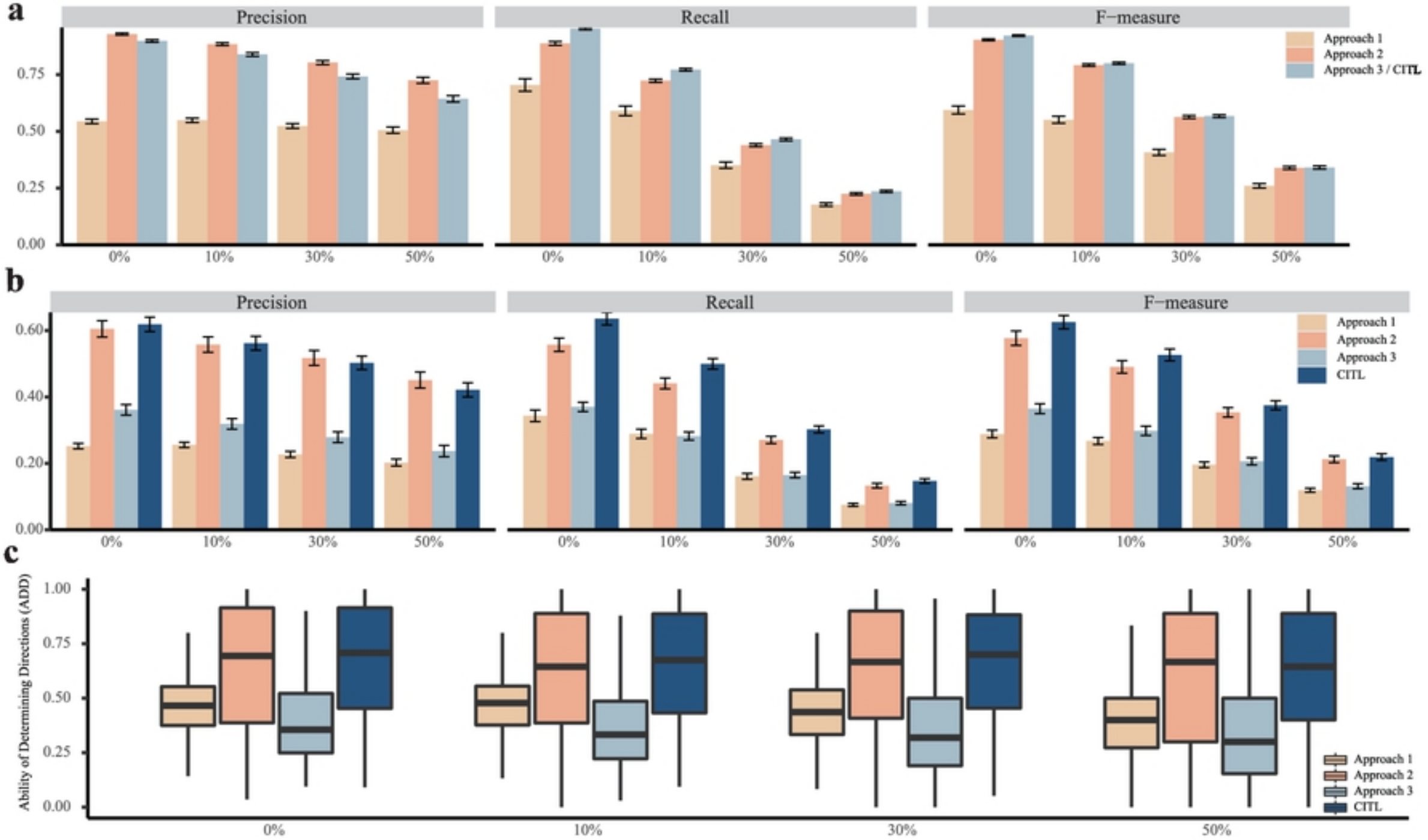
19. Tomás AR, Certal AC, Rodríguez-León J. FLRT3 as a key player on chick limb development. Dev Biol. 2011;355(2):324–33. doi:10.1016/j.ydbio.2011.04.031.

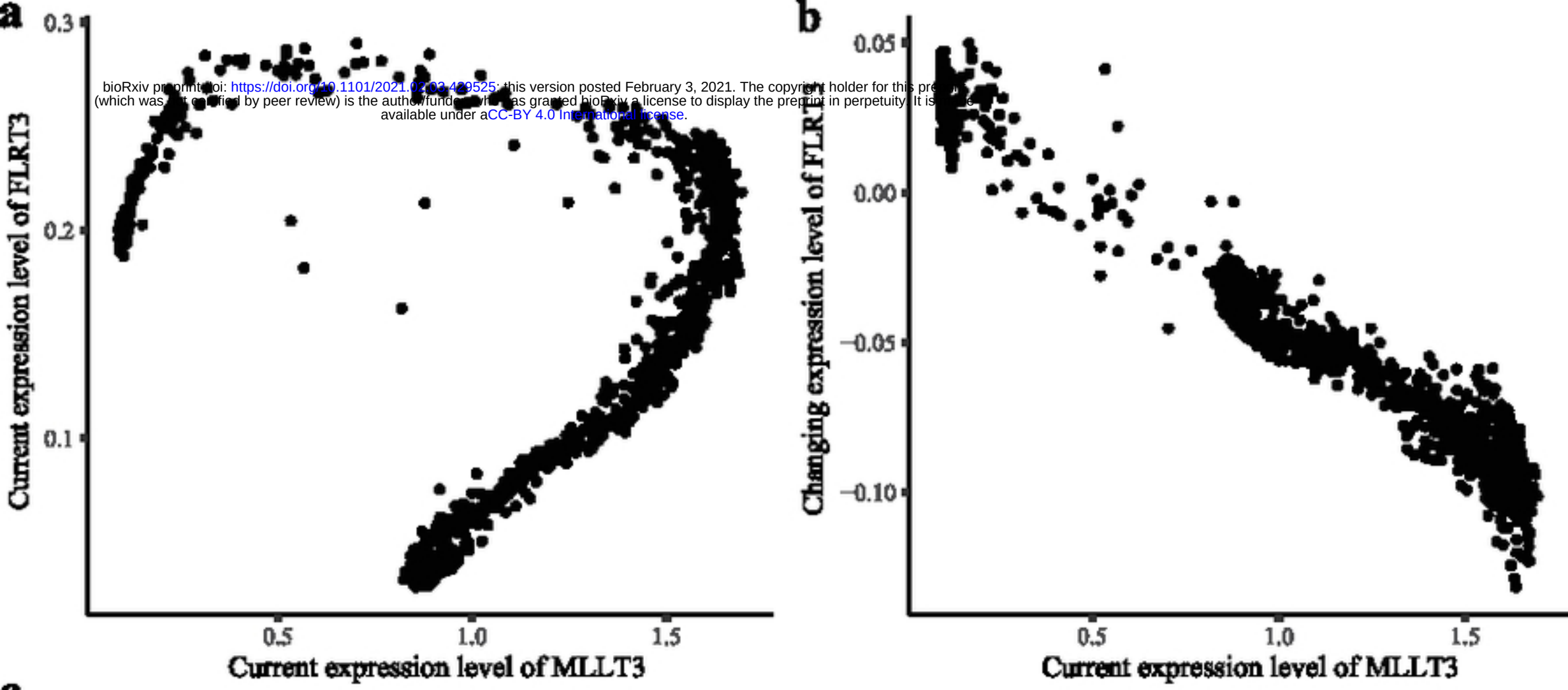
20. Martin M, Hua L, Wang B, Wei H, Prabhu L, Hartley AV, et al. Novel Serine 176 Phosphorylation of YBX1 Activates NF-κB in Colon Cancer. J Biol Chem. 2017;292(8):3433–3444. doi:10.1074/jbc.M116.740258.

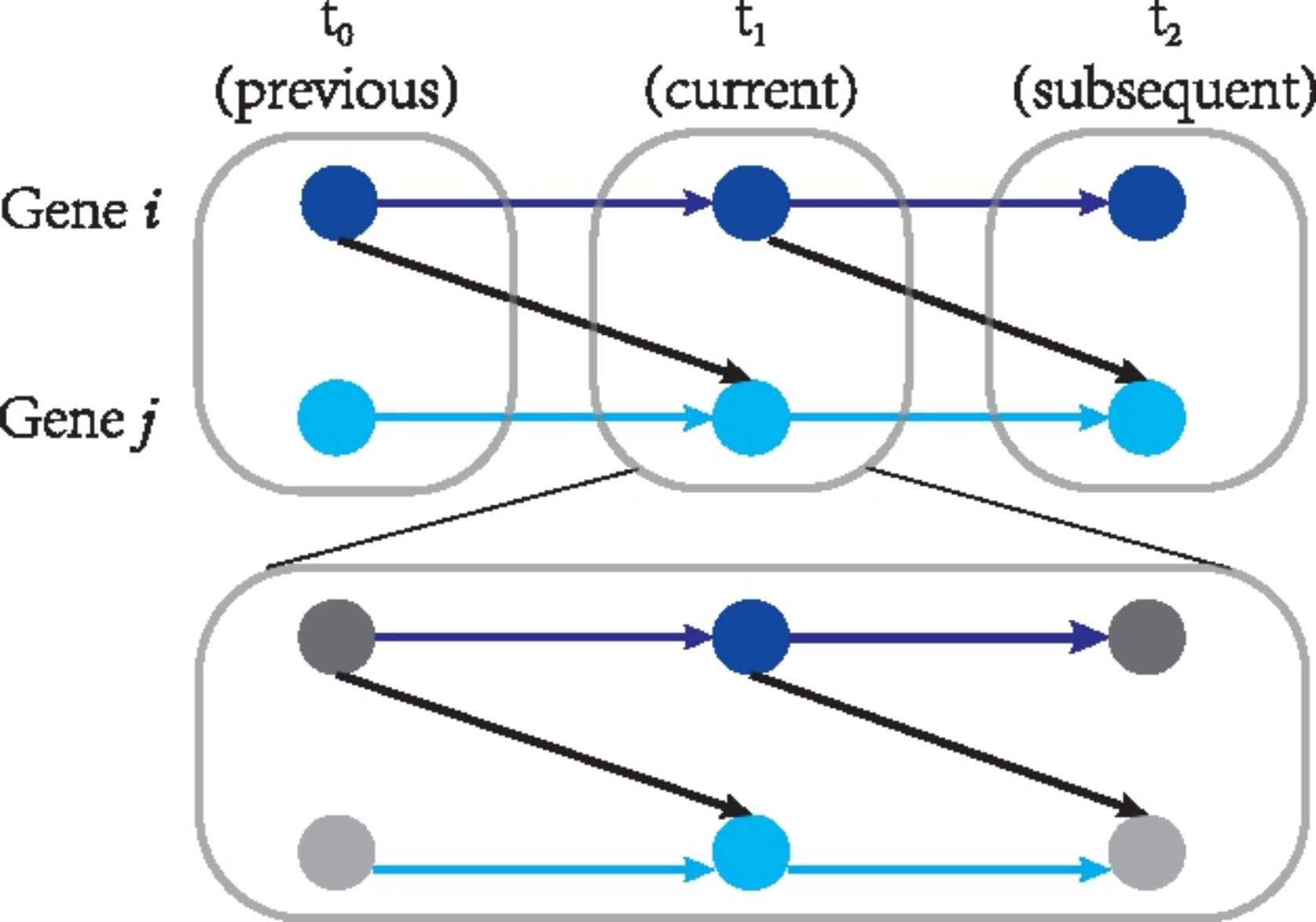
21. Zhang H, Zhou L, Zhou W, Xie X, Wu M, Chen Y, et al. EPS8-mediated regulation of multiple myeloma cell growth and survival. Am J Cancer Res. 2019;9(8):1622–1634.

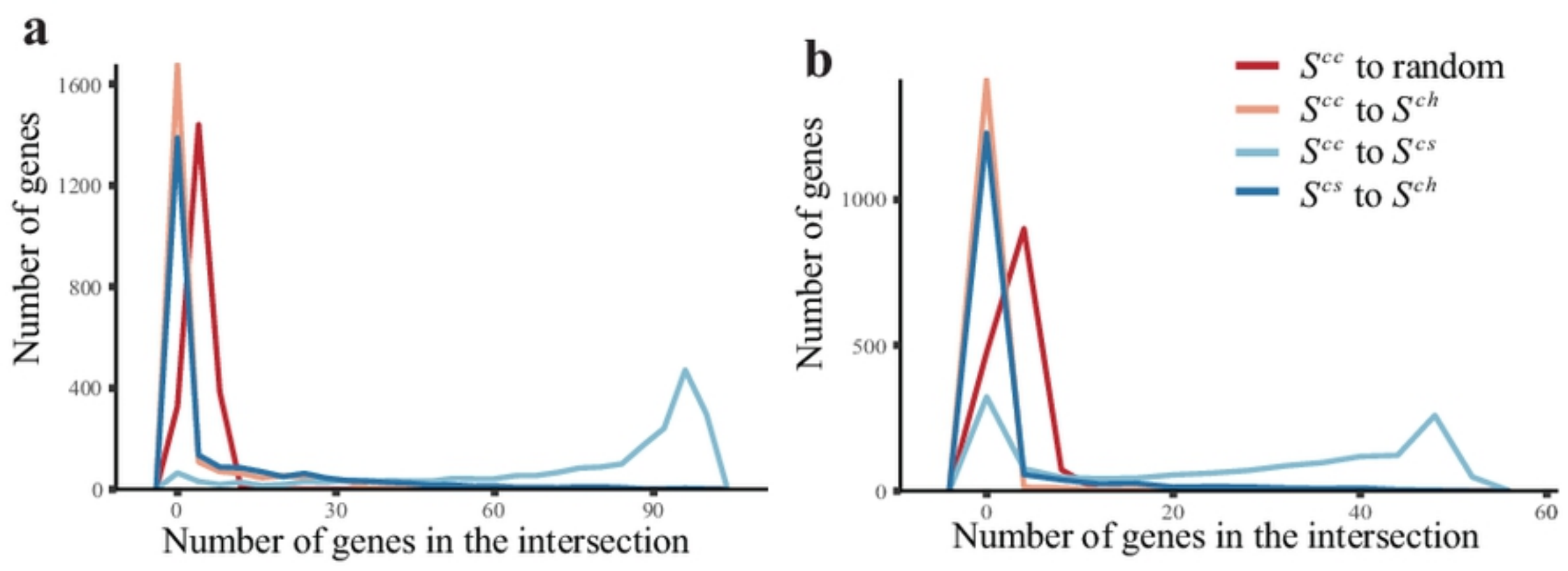
22. Krivanek J, Soldatov RA, Kastriti ME, Chontorotzea T, Herdina AN, Petersen J, et al. Dental cell type atlas reveals stem and differentiated cell types in mouse and human teeth. Nat Commun. 2020;11(1):4816. doi:10.1038/s41467-020-18512-7.

23. Dost AFM, Moye AL, Vedaie M, Tran LM, Fung E, Heinze D, et al. Organoids Model Transcriptional Hallmarks of Oncogenic KRAS Activation in Lung Epithelial Progenitor Cells. Cell Stem Cell. 2020;27(4):663 – 678.e8. doi:https://doi.org/10.1016/j.stem.2020.07.022.


a
Precision
Recall
F−measure
Approach 1
Approach 2
Approach 3 / CITL
0.00
0.25
0.50
0.75
0%
10%
30%
50%
b
Precision
Recall
F−measure
Approach 1
Approach 2
Approach 3
CITL
0.00
0.20
0.40
0.60
0%
10%
30%
50%
c
Ability of Determining Directions (ADD)
0.00
0.25
0.50
0.75
1.00
0%
10%
30%
50%
Approach 1
Approach 2
Approach 3
CITL




t_0
(previous)
t_1
(current)
t_2
(subsequent)
Gene *i*
Gene *j*


a
Number of genes
1600
1200
800
400
0
0
30
60
90
Number of genes in the intersection
b
Number of genes
1000
500
0
0
20
40
60
Number of genes in the intersection
S^{cc} to random
S^{cc} to S^{ch}
S^{cc} to S^{cs}
S^{cs} to S^{ch}