


## Systems biology

# Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference

Pierre-Cyril Aubin-Frankowski<sup>1,\*</sup> and Jean-Philippe Vert <sup>1,2,\*</sup>

<sup>1</sup>MINES ParisTech, PSL Research University, CBIO – Centre for Computational Biology, F-75006 Paris, France and <sup>2</sup>Google Research, Brain team, 75009 Paris, France

\*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on April 19, 2019; revised on May 4, 2020; editorial decision on May 27, 2020; accepted on June 11, 2020

## Abstract

**Motivation:** Single-cell RNA sequencing (scRNA-seq) offers new possibilities to infer gene regulatory network (GRNs) for biological processes involving a notion of time, such as cell differentiation or cell cycles. It also raises many challenges due to the destructive measurements inherent to the technology.

**Results:** In this work, we propose a new method named GRISLI for *de novo* GRN inference from scRNA-seq data. GRISLI infers a velocity vector field in the space of scRNA-seq data from profiles of individual cells, and models the dynamics of cell trajectories with a linear ordinary differential equation to reconstruct the underlying GRN with a sparse regression procedure. We show on real data that GRISLI outperforms a recently proposed state-of-the-art method for GRN reconstruction from scRNA-seq data.

**Availability and implementation:** The MATLAB code of GRISLI is available at: <https://github.com/PCAubin/GRISLI>.

**Contact:** pierre-cyril.aubin@mines-paristech.fr or jpvert@google.com

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables to observe genome-wide cellular activities at the single cell resolution (Kolodziejczyk *et al.*, 2015), generating extraordinary expectations for biologists and bringing forth new computational and mathematical challenges. By allowing us to study cell-to-cell variability, scRNA-seq has quickly become a technique of choice to systematically identify cell types in complex samples (Tasic *et al.*, 2016; Trapnell, 2015; Zeisel *et al.*, 2015) and understand dynamic biological processes, such as embryo development (Deng *et al.*, 2014), cell differentiation (Lönnerberg *et al.*, 2017) and cancer (Patel *et al.*, 2014).

A fascinating perspective offered by scRNA-seq studies is to understand how genes interact and regulate each other. In particular, by observing how gene expression varies among similar cells subject to stochastic fluctuations or involved in a dynamical process such as differentiation or cell cycle, one may be able to capture statistical or dynamical dependencies between genes which may, in turn, allow to reverse-engineer a gene regulatory network (GRN) to describe biologically which transcription factors (TFs) regulate which genes. While numerous algorithms have been proposed to infer GRN from bulk transcriptomic profiles (e.g. Marbach *et al.*, 2012, and references therein), scRNA-seq data raise new opportunities and challenges. On the one hand, the quantity of cells in scRNA-seq studies is often several-fold larger than the number of samples in bulk transcriptomic studies, offering increased statistical

power to capture regulatory interactions and allowing to capture subtle changes in dynamical process. On the other hand, scRNA-seq data are subject to various sources of variability (Kharchenko *et al.*, 2014; Risso *et al.*, 2018), and the precise type or state of each cell in a population must usually be inferred themselves from the data. In particular, in the case of dynamical processes, such as differentiation or cell cycles, several methods have been proposed to automatically infer a *pseudo-time* associated to each individual cell, as reviewed by Cannoodt *et al.* (2016).

As in bulk transcriptomic studies, putative functional interactions between genes can be detected by simple correlation analysis (Bacher and Kendzierski, 2016; Moignard *et al.*, 2013; Stegle *et al.*, 2015), or through more advanced strategies to capture statistical dependency between genes tailored to scRNA-seq data (Chan *et al.*, 2017; Filippi and Holmes, 2017). Aibar *et al.* (2017) refined the detection of gene modules by combining sequence information. However, such statistical associations do not necessarily capture regulatory relationships, which typically require perturbations or temporal experiments to be detected. As scRNA-seq studies can provide an ordering of cells involved in a dynamical process, through experimental time and/or inferred pseudo-time, they offer a unique opportunity to infer regulatory relationships by considering the (pseudo-)time information to compare gene expression profiles. For example, Herbach *et al.* (2017) propose a realistic, albeit complex, stochastic dynamical system to model scRNA-seq data, which is only tested on simulated data for networks of two genes due to its

computational complexity. Moignard *et al.* (2015) presented a formalism to infer a Boolean network from single-cell qRT-PCR data but requires to discretize gene expression values to an on/off status in each cell. Ocone *et al.* (2015) proposed to infer a GRN by estimating an ordinary differential equation (ODE) from pseudo-time-ordered scRNA-seq data; however, due to the computational complexity of the model selection procedure, the final GRN is limited to be a refinement of a coarse GRN inferred with GENIE3 (Huynh-Thu *et al.*, 2010), a method for bulk gene expression. A similar, linear ODE-based formalism was proposed by Matsumoto *et al.* (2017), who designed a more efficient procedure named SCODE to directly infer *de novo* the GRN from scRNA-seq data. SCODE assumes that all cells are on the same trajectory, and estimates the parameters of the ODE by integrating it and optimizing the fit between the integrated model and each individual cell's transcriptome. However, the resulting optimization problem is computationally intractable and is solved only approximately by restricting the class of GRN models.

In this work, we follow the same linear ODE-based formalism as SCODE for GRN inference from scRNA-seq data, and propose a new approach, which we name GRISLI, to estimate the parameters of the model (Supplementary Fig. S1). GRISLI first estimates the *velocity* of each cell, i.e. how each gene's expression is increasing or decreasing in the dynamical process for each cell, and then estimates the structure of the GRN by solving a sparse regression problem to relate the gene expression of a cell to its velocity profile. We solve the sparse regression problem with a variant of stability selection (Meinshausen and Bühlmann, 2010) proposed in TIGRESS (Haury *et al.*, 2012), a method for GRN inference from bulk transcriptomics where no velocity is involved since samples are assumed to be near steady state. In spite of a similar ODE formalism, GRISLI differs from SCODE in several aspects: (i) while SCODE assumes that all cells are on the same trajectory, in GRISLI, we consider *bundles* of trajectories derived from a large number of initial conditions and do not integrate the ODE; (ii) while SCODE integrates the ODE, leading to a computationally intractable optimization problem to infer the parameters, we solve a simple, convex regression problem that allows us to make no restrictive assumption on the GRN structure and leads to a fast algorithm. These benefits come at the cost of estimating the velocity of each cell, for which we propose a novel procedure based on weighted averages of finite differences with other cells at nearby positions in space–time.

We empirically assess the performance of GRISLI on human and murine scRNA-seq data and show that it outperforms TIGRESS, highlighting the benefits of the ODE-based framework for scRNA-seq data, as well as outperforming SCODE, confirming the relevance of our new estimation procedure.

## 2 Materials and methods

### 2.1 Setting and notations

We consider the problem of inferring a GRN from a set of  $C$  single-cell transcriptomic profiles  $x_1, \dots, x_C \in \mathbb{R}^G$ , where  $x_i \in \mathbb{R}^G$  represents the expression, for the  $i$ th cell, of  $G$  genes. We furthermore assume that the cells are involved in a dynamical process, such as differentiation or cell cycle, and that for each cell  $i \in [1, C]$ , we have an estimate of a time-label  $t_i \in \mathbb{R}$  that describes where the cell is in the process. The time-label  $t_i$  is assigned to the  $i$ th cell based either on the real experimental time or on a calculated pseudo-time. Hence, we assume given a collection of time-labeled vectors  $\{(x_i, t_i) \in \mathbb{R}^G \times \mathbb{R} : i = 1, \dots, C\}$ .

We model the dynamical process of the cell expression  $x(t) \in \mathbb{R}^G$  as a linear ODE of the form

$$\frac{dx}{dt} = Ax, \quad (1)$$

where  $A \in \mathbb{R}^{G \times G}$  characterizes how each gene's expression level influences the expression dynamics of other genes. While linear ODE is simplified models of gene regulation, we show in Supplementary Material (Supplementary Figs S2 and S3) that they

are good enough to capture significant amounts of variation in real data. Assuming that, each gene is regulated by only a few TFs, we assume that  $A$  is sparse, in the sense that  $A_{ij} \neq 0$  means that the expression of gene  $j$  influences that of gene  $i$ , i.e. that gene  $j$  regulates gene  $i$ .

Inferring the GRN thus amounts to estimating which entries in  $A$  are non-zero. To do so, we propose a two-step approach called GRISLI: first, we estimate the velocity of each cell  $v_i = dx_i/dt$  with an estimator  $\hat{v}_i$ , and second, we infer non-zero elements of  $A$  by estimating the support of the regression model (1) from the sample  $(x_i, \hat{v}_i)_{i=1, \dots, C}$  with a stability selection procedure. We detail each step in turn below.

### 2.2 Velocity inference

Given the set of time-labeled vectors  $\{(x_i, t_i) \in \mathbb{R}^G \times \mathbb{R} : i = 1, \dots, C\}$ , we estimate the velocity  $\hat{v}_i \in \mathbb{R}^G$  of each cell  $i \in [1, C]$  as follows. We first observe that from any other cell  $(x_j, t_j)$ , with  $t_j \neq t_i$ , we may form the following velocity estimate based on finite difference:

$$\hat{v}_{ij} = \frac{x_j - x_i}{t_j - t_i}.$$

This estimate is interesting only when (i)  $t_j$  is not too far from  $t_i$ , so that a finite difference is a good approximation of the derivative, and (ii) the trajectories of cells  $j$  and  $i$  are close to each other, in the sense that if we were able to observe cell  $j$  at time  $t_i$  then it should be close to  $x_i$ . Based on these considerations, we form the velocity estimate  $\hat{v}_i$  as a weighted average of the  $\hat{v}_{ij}$ , with weights defined by a spatio-temporal kernel  $K(x, t, x', t')$  that quantifies how we believe  $(x', t')$  is useful to estimate the velocity at  $(x, t)$ . However, as the points living in the past or the future of a given point  $(x_i, t_i)$  act differently on the velocity, we separate their contributions into two weighted averages.

$$\hat{v}_i = \frac{1}{2} \frac{\sum_{j|t_j > t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{ij}}{\sum_{j|t_j > t_i} K(x_i, t_i, x_j, t_j)} + \frac{1}{2} \frac{\sum_{j|t_j < t_i} K(x_i, t_i, x_j, t_j) \hat{v}_{ji}}{\sum_{j|t_j < t_i} K(x_i, t_i, x_j, t_j)}.$$

As for the spatio-temporal kernel  $K$ , we arbitrarily take the following:

$$\begin{aligned} \forall (x, t, x', t') \in \mathbb{R}^G \times \mathbb{R} \times \mathbb{R}^G \times \mathbb{R}, \\ K(x, t, x', t') = (t - t')^2 \exp\left(-\frac{(t - t')^2}{2\sigma_t^2}\right) \\ \times \exp\left(-\frac{\|x - x'\|_{\mathbb{R}^G}^2}{2\sigma_x^2}\right), \end{aligned}$$

where  $\sigma_x$  and  $\sigma_t$  are fixed, respectively, to the square root of the 10th percentile of the distribution of squared distances in space,  $x$  (respectively in time,  $t$ ).

### 2.3 GRN inference

Once we form the estimate  $\hat{v}_i$  for the velocity  $v_i = dx_i/dt$  of each cell  $i = 1, \dots, C$ , we estimate the GRN by considering (1) as a sparse regression problem of the form  $\hat{v} = Ax$ , with observations  $(x_i, \hat{v}_i)_{i=1, \dots, C}$ . We estimate the non-zero entries of  $A$  using a stability selection procedure for sparse regression (Haury *et al.*, 2012; Meinshausen and Bühlmann, 2010). More precisely, for each candidate regulator  $j \in [1, C]$  and target gene  $i \in [1, C]$ , we compute a score  $s(i, j) \in (0, 1)$  which increases when we believe that  $A_{ij} \neq 0$ , i.e. that  $j$  regulates  $i$ . The score  $s(i, j)$  itself depends on three parameters  $R, L \in \mathbb{N}$  and  $\alpha \in [0, 1]$ , and is computed through the procedure described by Haury *et al.* (2012), which we now summarize. Let us denote by  $X := (x_1, \dots, x_C) \in \mathbb{R}^{G \times C}$  the expression matrix and  $\hat{V} := (\hat{v}_1, \dots, \hat{v}_C) \in \mathbb{R}^{G \times C}$  the matrix of estimated velocities. We repeat  $R$  times a procedure where we create a new expression matrix  $\tilde{X}$  and a new velocity matrix  $\tilde{V}$  obtained from  $X$  and  $\hat{V}$  by (i) randomly sub-sampling  $\lfloor C/2 \rfloor$  columns (i.e. cells) simultaneously from  $X$  and  $\hat{V}$ ,

and (ii) multiplying each row  $i$  of  $\mathbf{X}$  by a different random number  $\beta_i$  uniformly sampled between  $\alpha$  and 1. We then estimate for every  $(\tilde{\mathbf{X}}, \tilde{\mathbf{V}})$  a sparse matrix  $\mathbf{A}$  by solving a lasso regression problem:

$$\min_{\mathbf{A} \in \mathbb{R}^{G \times G}} \|\tilde{\mathbf{V}} - \mathbf{A}\tilde{\mathbf{X}}\|_2^2 + \lambda \|\mathbf{A}\|_1, \quad (2)$$

over a grid of regularization parameters  $\lambda$  ensuring that we have solutions having from 0 to at least  $L$  non-zero entries in each row of  $\mathbf{A}$ . For each pair of genes  $(i, j)$  and each integer  $l \in [1, L]$ , we measure, among the  $R$  repeats, the frequency  $F(i, j, l)$  at which  $\mathbf{A}_{ij}$  is non-zero for the solution of (2) when  $\lambda$  is set such that  $l$  entries are non-zero in the  $i$ th row of  $\mathbf{A}$ , i.e. when the  $j$ th TF is among the top  $l$  TFs in the regularization path to explain the expression of the  $i$ th gene. We then consider the *area score* proposed by Haury et al. (2012):

$$s_{\text{area}}(i, j) = \frac{1}{L} \sum_{l=1}^L F(i, j, l).$$

Alternatively, we can consider the original stability selection score proposed by Meinshausen and Bühlmann (2010):

$$s_{\text{original}}(i, j) = F(i, j, L).$$

Haury et al. (2012) discusses the differences between both scores, and suggests to prefer the area score which is therefore our preferred choice.

The choice of the three parameters  $R$ ,  $L$  and  $\alpha$  is a difficult question. While  $R$  should typically be as large as possible to reduce random fluctuations of the algorithm,  $\alpha$  should typically be chosen in the range  $[0.2, 0.8]$  according to Haury et al. (2012), and  $L$  should be tested on a large grid of values. In the experiments below, we provide results for different values of these parameters to demonstrate the potential of the method. In other applications where some interactions are known, we suggest as well to test predictions over a grid of values, and to pick the model that best matches known interactions.

## 2.4 Data

To test GRISLI, we first use the two datasets provided online by Matsumoto et al. (2017), which we refer to as the *Matsumoto benchmark*. These are scRNA-seq datasets where the TF expression is considered as the log-transform of the transcripts per million reads or the log-transform of the fragments per millions of kilobases mapped. The first dataset, published by Treutlein et al. (2016) (named *Data2* in SCODE) comes from a direct reprogramming of murine embryonic fibroblast cells to myocytes at days 0, 2, 5 and 22. This dataset contained 373 cells. The second dataset (named *Data3* in SCODE) comes from Chu et al. (2016) and measures the differentiation of human ES cells to definitive endoderm cells, taken at 0, 12, 24, 36, 72 and 96 h. This dataset contained 758 cells. Following Matsumoto et al. (2017), we keep only the 100 TF with largest variance in each dataset.

We also consider a dataset of 3696 murine pancreatic cells proposed by Bastidas-Ponce et al. (2019) to study pancreatic endocrinogenesis (GEO accession number GSE132188). We obtained and processed the data following the default pipeline of scvelo software (Bergen et al., 2019), keeping only the 179 TF with at least 20 counts in unspliced and spliced reads, and log-transforming the counts.

The list of TF and gold standard regulatory network comes from the RIKEN mouse TFDB for mouse (Kanamori et al., 2004) and animalTFDB for human (Hu et al., 2018), which contain experimentally derived TF occupancy sites based on genomic DNase I footprinting (Neph et al., 2012; Stergachis et al., 2014). We downloaded the data from the Transcription Factor Regulatory Network database (<http://www.regulatorynetworks.org>), merging together all interactions found across different cell types.

## 2.5 Performance evaluation

We evaluate the performance of each GRN inference method by its area under the receiver operating characteristic curve (AUC), calculated by comparing the predictions (a score for each pair of TFs) to the gold standard regulatory networks. For the sake of comparison, we follow the choices made by Matsumoto et al. (2017): we compute the AUC ignoring self-loops (the diagonal elements) and excluding the TFs (among the 100 TFs selected as having the largest variance) that do not have an edge in the true network, resulting in a subnetwork of 40 TFs for the murine dataset and 49 TFs for the human dataset.

## 2.6 Other GRN inference methods

In addition to GRISLI, SCODE and TIGRESS, we test two baseline methods: a mutual information (MI) network, using the infotheo R package run with the ‘mm’ parameter, and a causal Bayesian network (Maathuis et al. 2010) using the idafast function of the pcalg R package. In both cases, the ROC curve is computed based on the absolute values of the weights given to the edges of the GRN. As mutual information is symmetric, we also symmetrized the ground truth for fair comparison. IDA computes a causality directed acyclic graph and sometimes shows several weights values for a given edge, but never more than two in our case [we then pick the one with smallest absolute weight (Choosing the mean of the weights led to very similar AUROC values.)].

## 2.7 Other velocity inference

Besides our velocity inference method described in Section 2.2, we test scvelo (Bergen et al., 2019), a method generalizing the pioneering work of La Manno et al. (2018) to infer velocities based on modelling the relative number of unspliced and spliced reads for each gene. To estimate velocities on the pancreatic cell dataset, we follow precisely the corresponding tutorial of scvelo (<https://scvelo-notebooks.readthedocs.io/Pancreas.html>).

## 3 Results

### 3.1 GRISLI compared to SCODE

We propose a new method for Gene Regulatory network Inference from scRNA-seq data with Linear differential equations (GRISLI). The input to GRISLI is a set of time-stamped scRNA-seq data  $(x_i, t_i)_{i=1, \dots, C}$ , where  $C$  is the number of cells,  $x_i$  is the vector of gene expression for the  $i$ th cell and  $t_i$  is the time associated to the  $i$ th cell; this time can be based either on the real experimental time, or on a calculated pseudo-time. GRISLI combines the dynamical model of SCODE (Matsumoto et al., 2017) with the statistical procedure for network estimation of TIGRESS (Haury et al., 2012). More precisely, like SCODE, we model the dynamics of gene expression as a linear differential equation  $dx/dt = \mathbf{A}x$ , where  $\mathbf{A}$  is sparse and encodes the GRN in its non-zero entries. By integrating this equation, and assuming all cells have the same (unknown) initial stage  $x_0$ , Matsumoto et al. (2017) proposed to estimate  $\mathbf{A}$  by solving:

$$\min_{\mathbf{A} \in \mathbb{R}^{G \times G}} \sum_{i=1}^C \|x_i - e^{t_i \mathbf{A}} x_0\|^2, \quad (3)$$

which is, however, a computationally intractable non-convex optimization problem in  $\mathbf{A}$ ; to overcome the difficulty, Matsumoto et al. (2017) restricted themselves to low-rank diagonalizable matrices of the form  $\mathbf{A} = \mathbf{W}\mathbf{B}\mathbf{W}^+$ , where  $\mathbf{B}$  is diagonal of small rank, and use further assumptions and heuristics to obtain a tractable algorithm SCODE to optimize successively  $\mathbf{W}$  and  $\mathbf{B}$ .

GRISLI is based on the same dynamical model as SCODE, but exploits it differently. Instead of integrating the dynamical model  $dx/dt = \mathbf{A}x$  as in Equation (3), we see it as a regression problem of the form  $v = \mathbf{A}x$  where  $v = dx/dt$  is the velocity of each cell and take a two-step approach to estimate  $\mathbf{A}$ : (i) first estimate the instant velocity  $\hat{v}_i$  of each cell  $i = 1, \dots, C$ , and (ii) then estimate the non-zero entries of  $\mathbf{A}$  using a stability selection procedure akin to the one

used in TIGRESS to identify non-zero coefficients in the regression problem  $\hat{v} = Ax$  from samples  $(x_i, \hat{v}_i)_{i=1, \dots, C}$ . The technical details of GRISLI are presented in Section 2.

While GRISLI involves a step of velocity inference absent from SCODE, the benefits of the GRISLI model over the SCODE model include the facts that (i) we do not need to assume that all cells lie on the same trajectory, and (ii) we make no restricting assumption on  $A$ , such as being of low rank and diagonalizable, and still derive a computationally efficient convex problem to estimate  $A$ .

### 3.2 Performance on the Matsumoto benchmark

To assess the predictive capacity and the speed of GRISLI, we test it on two benchmark datasets analyzed by Matsumoto *et al.* (2017): (i) a murine dataset of 373 cells corresponding to direct reprogramming of murine embryonic fibroblast cells to myocytes at days 0, 2, 5 and 22 (Treutlein *et al.*, 2016), and a human dataset of 758 cells corresponding to differentiation of human ES cells to definitive endoderm cells, taken at 0, 12, 24, 36, 72 and 96 h (Chu *et al.*, 2016). We follow the experimental protocol of Matsumoto *et al.* (2017) to assess the effectiveness of GRISLI to predict known regulations, and compare it to SCODE and TIGRESS. All methods having a stochastic component, we run them 30 times on each dataset and summarize their performance by the distribution of AUC scores (see Section 2). The AUC takes values between 0.5 for a random prediction to 1 for a perfect recovery of known regulations.

Figure 1 summarizes the performance of the three methods on both datasets, and Supplementary Figure S4 shows the corresponding ROC and precision-recall curves. Since each method depends on several parameters, we tested different parameter sets, as detailed in the next section, and report here the best performance of each method to assess how well they perform if we choose good parameters. The parameters used for SCODE were the one provided for their respective datasets in the study by Matsumoto *et al.* (2017). The parameters used for GRISLI result from an extensive search, illustrated in Figure 2. The parameters used for TIGRESS were obtained similarly, in practice, they coincide with the parameters of GRISLI.

We first notice that, in both cases, TIGRESS has the poorest performance, which highlights the limitations of GRN techniques

developed for bulk RNA-seq in the context of scRNA-seq data. This is coherent with the findings of Matsumoto *et al.* (2017) who noticed that several state-of-the-art method for GRN inference from bulk RNA-seq data performs poorly on these scRNA-seq data. To further illustrate this, we report in Table 1, the mean AUC on both datasets of GRISLI, SCODE, TIGRESS and seven other methods for GRN inference: a MI network, a causal discovery algorithm based on Bayesian networks (IDA), a correlation network (Cor), a regression network based either on normal linear regression (lm), or on lasso sparse regression (mspgs), a regression-tree-based approach (GENIE3) and its extension to time-course expression data (JUMP3). The results for the last five methods are taken from the study by Matsumoto *et al.* (2017). These results show that, on this benchmark, GRISLI and SCODE clearly outperform all other methods, suggesting that the explicit dynamical model they use is beneficial for GRN inference from scRNA-seq data.

Second, we see that GRISLI outperforms SCODE on both datasets: on the murine benchmark, GRISLI has a mean AUC of 0.571 versus 0.528 for SCODE, while on the human benchmark GRISLI reaches an AUC of 0.573 versus 0.550 for SCODE. Contrary to Matsumoto *et al.* (2017), we did not provide only one AUC value to assess the performance but a boxplot that shows the inherent stochasticity of the methods. As both SCODE and GRISLI share the same underlying dynamical model, this highlights the benefits of the GRISLI approach to estimate the parameters of the system. We furthermore notice that the variability in the performance across runs is smaller for GRISLI than for SCODE. In terms of computational cost, GRISLI and TIGRESS ran faster (respectively, 20 and 100 s on the human and murine datasets, on a 4-cores 3.5 GHz Intel Core i5 with 16GB of 667 MHz DDR3 RAM) than SCODE (500 s on both datasets).

For the murine dataset, we used Monocle pseudo-time as a time-label. However, for the human dataset, we used the real experimental time values, as the pseudo-time AUC results did not increase with  $R$ . It is possible that when the measurements are close enough and evenly spaced in time (such as for the human data), real time is sufficient, whereas for more distant experiences (which is the case of the murine data), the pseudo-time may be of some interest.

We provide in Supplementary Tables S1 and S2, the top regulation predictions by GRISLI on the murine and human datasets,

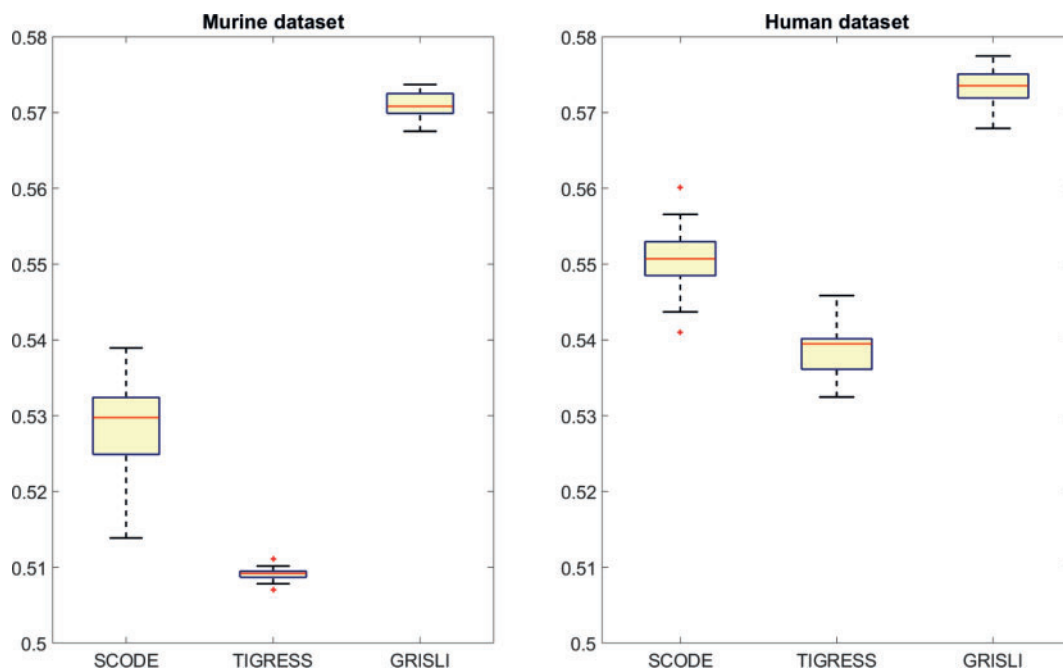


Fig. 1. Performance of the different methods (as distribution of AUC over 30 repeats) on the murine (left) and human (right) benchmarks. SCODE score was obtained taking the average of 50 replicates with the rank  $D$  equal to 4 and 100 trials. GRISLI has, respectively, as parameters  $L = 70$ ,  $R = 1500$  and  $\alpha = 0.3$  for the murine benchmark,  $L = 1$ ,  $R = 3000$  and  $\alpha = 0.3$  for the human benchmark. TIGRESS was run with the same parameters as GRISLI



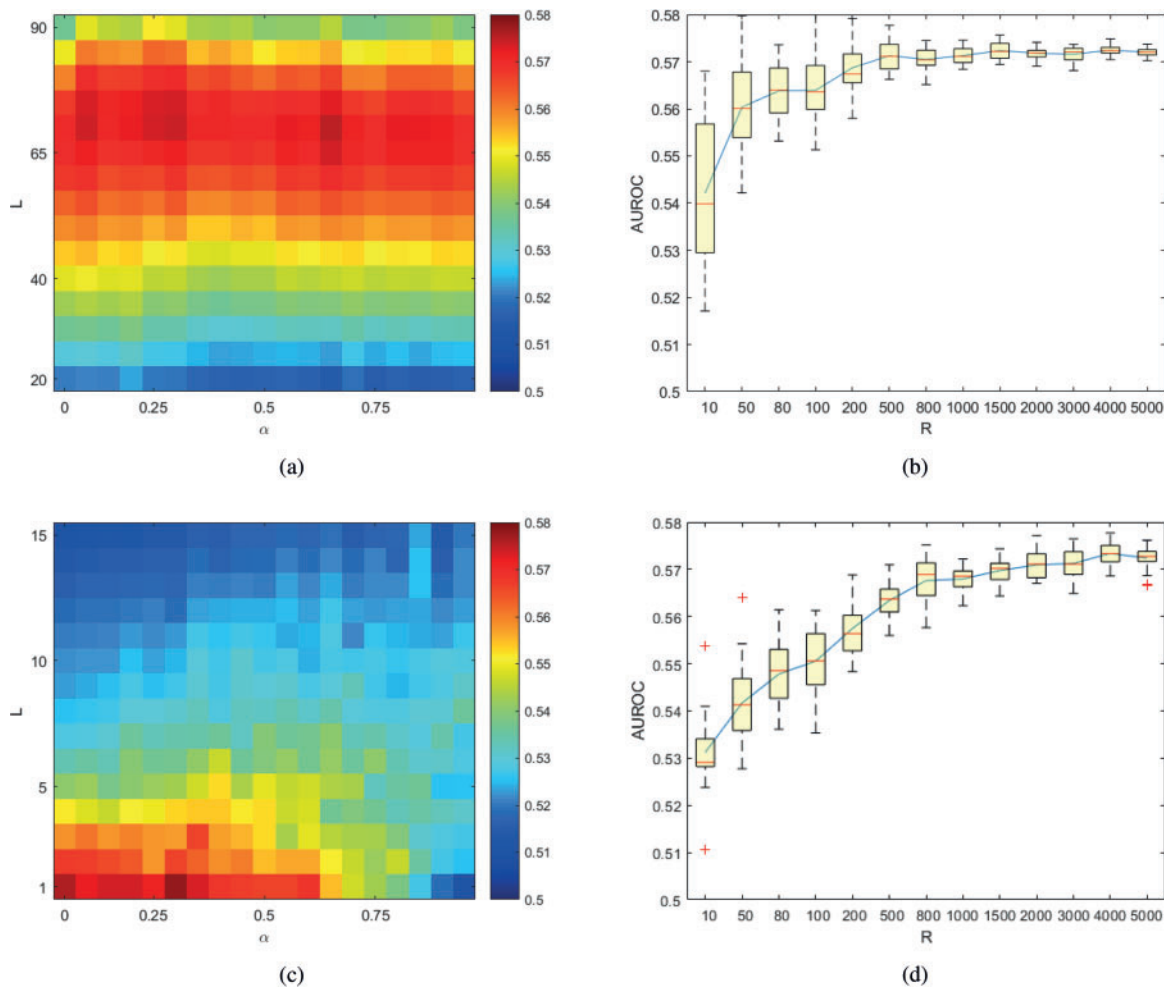


Fig. 2. (a) Performance of GRISLI (AUC) on the murine dataset with  $R = 1500$  and varying  $\alpha$  and  $L$ . (b) Performance of GRISLI (AUC) on the murine dataset with  $\alpha = 0.4$ ,  $L = 70$  and varying  $R$  (repeated 20 times). (c) Same as (a) but for the human data and  $R = 3000$ . (d) Same as (b) but for the human data with  $\alpha = 0.3$  and  $L = 1$

**Table 1.** Performance (mean AUC) of different GRN inference methods on the Matsumoto benchmark

|                | GRISLI       | SCODE | TIGRESS | MI    | IDA   | Cor   | lm    | msgps | GENIE3 | JUMP3 |
|----------------|--------------|-------|---------|-------|-------|-------|-------|-------|--------|-------|
| Murine dataset | <b>0.571</b> | 0.528 | 0.509   | 0.448 | 0.463 | 0.492 | 0.489 | 0.516 | 0.472  | 0.492 |
| Human dataset  | <b>0.573</b> | 0.551 | 0.539   | 0.490 | 0.502 | 0.524 | 0.480 | 0.499 | 0.522  | 0.501 |

Note: Results for Cor, lm, msgps, GENIE3 and JUMP3 are taken from [Matsumoto et al. \(2017\)](#). The values in bold show, for each dataset, the best performance across methods.

respectively. While no discernible pattern emerges among the targets, we see that GRISLI predicts as top regulators Stat1 for the murine data and NANOG and GATA6 for the human data, which are known to be important for the differentiation of human ES cells to fibroblasts ([Bessonnard et al., 2014](#)).

### 3.3 Sensitivity to the parameters

Here, we investigate in more details the influence of the parameters  $L$ ,  $\alpha$  and  $R$  on the performance of TIGRESS. While  $R$  should typically be chosen as large as possible to ensure that the empirical average converges to the expectation in the procedure, the optimal choice of  $L$  and  $\alpha$  is harder to predict.

We therefore systematically assess the performance of GRISLI, in terms of AUC, over a large grid of values for  $L$ ,  $\alpha$  and  $R$ . [Figure 2](#) summarizes the results, for both the murine and the human datasets. As expected, increasing  $R$  is always beneficial. For example,

[Figure 2b and d](#) shows for a particular choice of  $L$  and  $\alpha$  that, after swiftly increasing, the AUC values reaches a plateau when  $R$  increases. As  $R$  has a significant effect on runtime, we take an intermediate value of  $R = 1500$  for the murine dataset and  $R = 3000$  for the human dataset for the subsequent experiments.

The influence of  $L$  and  $\alpha$  is, as expected, more complex and depends on the dataset. On the murine dataset, the AUC is fairly stable and optimal for any  $L \in [62, 78]$  with little influence of  $\alpha$  ([Fig. 2a](#)), whereas on the human dataset only the value  $L = 1$  seems adequate,  $\alpha$  having a bigger effect as the scores sharply diminish for  $\alpha$  larger than 0.6 ([Fig. 2c](#)). Nonetheless, the overall shape is coherent with the analysis of [Hauray et al. \(2012\)](#): there is a compensating effect, as a smaller  $\alpha$  increases the diversity between the batches. On the contrary, reducing  $L$  limits the number of selected edges of the regulation network, making the predictions more similar.

The optimal values for  $L$  are strikingly different between the two datasets. While taking  $\alpha$  equal to 0.3 or  $R$  larger than 1000 seem

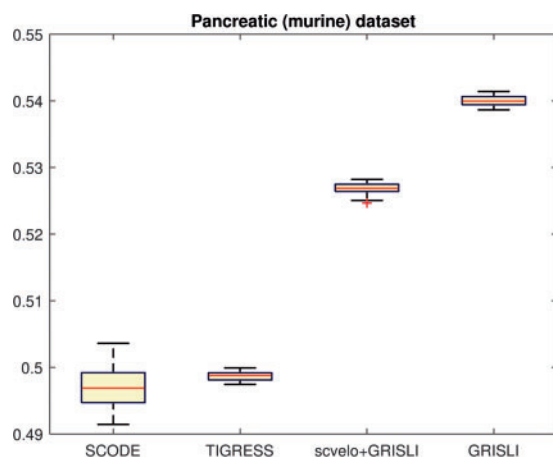


Fig. 3. Performance of the different methods (as distribution of AUC over 30 repeats) on the pancreatic data. SCODE score was obtained taking the average of 10 replicates with the rank  $D$  equal to 4 and 20 trials. GRISLI has parameters  $L=100$ ,  $R=200$  and  $\alpha=0.4$ . TIGRESS was run with the same parameters as GRISLI. We compared also the case of replacing the velocity considered in GRISLI by the output of *scvelo*

adequate in both cases, determining heuristics to choose  $L$  is still an open problem. We found experimentally that the best value of  $L$  seems related to the expected sparsity of the predicted network, defined here as the proportion of active edges of the gold standards, which is of 7% of the murine data, and of 4% of the human data. This value reflects the connectivity of the network. In practice, we suggest to test different values of  $L$  and select the one that best recapitulates known interactions, if any.

### 3.4 Influence of velocity inference

The first step of GRISLI, the velocity inference, compares the expression profile of each cell to the expression profiles of similar cells as explained in Section 2.2. In contrast, La Manno *et al.* (2018) proposed a completely different methodology to infer velocities of single cells, based on a model for the relative quantities of unspliced and spliced reads for each transcript. The method was later generalized and improved by Bergen *et al.* (2019) in the *scvelo* software. Since velocity inference and network inference are two independent steps in GRISLI, we now investigate the performance of GRISLI using different velocity inference methods.

For that purpose, we consider the pancreatic endocrinogenesis dataset, which is used as tutorial by Bergen *et al.* (2019) to demonstrate the relevance of velocity inference by *scvelo*. Following the corresponding tutorial, we extract the expression matrix of 179 TF for 3696 cells, together with a pseudo-time estimate for each cell and a velocity matrix estimated by *scvelo*. In parallel, we estimate a second velocity matrix through the GRISLI methodology described in Section 2.2, using the expression matrix and pseudo-time only. We then assess the performance of GRISLI with each of the velocity matrices and also test SCODE and TIGRESS on this dataset. Figure 3 shows the AUC values of the various methods on this new dataset, and Supplementary Figure S5 shows in detail the ROC and precision-recall curves. We observe that SCODE and TIGRESS have very poor performance, whereas GRISLI significantly outperforms them. Interestingly, we see that GRISLI's velocity inference method leads to slightly better results than *scvelo*'s, suggesting that the method used to infer velocities has some impact on the quality of the final GRN inferred, and that GRISLI's approach for velocity inference is competitive with other existing methods.

## 4 Discussion and conclusion

Based on the (pseudo-)time information of scRNA-seq data, we propose GRISLI, a new method to infer GRN without any other

information than the scRNA-seq data themselves. GRISLI is based on the same linear ODE formalism as SCODE, where the GRN is defined as the support of the matrix that relates TF expression to velocities of other genes; however, GRISLI differs from SCODE in several assumptions. While SCODE assumes that all cells are on the same trajectory, which allows to model each cell by integrating the ODE from a unique initial condition, GRISLI considers bundles of trajectories where each cell may be following a unique trajectory, governed by a unique ODE common to all cells. Given the inherent stochasticity in gene expression, and the well-known bifurcations possible when similar cells differentiate in different subtypes (Paul *et al.*, 2015), we believe that allowing cells to evolve on different trajectories is an important property. This flexibility prevents us from integrating the ODE as in SCODE, and forces us instead to estimate the local velocity of each cell. We propose a simple estimator based on a weighted average of finite differences between pairs of cells, and believe that much work remains to be done for velocity inference from scRNA-seq data. Interestingly, La Manno *et al.* (2018) and Bergen *et al.* (2019) proposed recently completely different approaches for velocity inference, by comparing the quantity of spliced versus unspliced mRNA; as we have seen, our estimator slightly outperforms these other approaches in terms of GRN inference accuracy, which suggests that further improvement in the velocity inference step of GRISLI may translate into more accurate GRN inference. Another difference between GRISLI and SCODE concerns the assumptions on the structure of the GRN. For computational reasons, SCODE constrains the GRN matrix to be low-rank, while GRISLI puts no assumption on it. Furthermore, if we wish to constrain the structure of the GRN in GRISLI, it can easily be achieved by adding structured sparsity constraints in the sparse regression problem (Bach *et al.*, 2012).

In terms of performance, we observed that GRISLI outperforms both SCODE and TIGRESS on both human and murine scRNA-seq data. The limited performance of TIGRESS highlights the fact that methods developed for bulk RNA-seq data, based on the assumption that samples are near a steady-state condition, are not optimal for single-cell data. This was already observed by Matsumoto *et al.* (2017) with other state-of-the-art GRN inference methods for bulk RNA-seq, and confirms the relevance of interpreting the GRN as the support of the matrix that relates expression to velocity in the linear ODE framework. The fact that GRISLI outperforms SCODE, in contrast, confirms the relevance of our assumptions and estimation procedure. However, we should keep in mind that the performance in absolute value remains modest, with an average AUC of 0.575. This is roughly similar to the best performances reached on simpler organisms from bulk transcriptomic data (Marbach *et al.*, 2012), and highlights again the difficulty of *de novo* GRN inference. Several reasons may be responsible for the difficulty to reverse engineer GRN from transcriptomic data. First, the concentration of messenger RNA of a TF is not always a reliable estimate of the active protein level, although, as discussed in the study by Matsumoto *et al.* (2017), this problem may be mitigated to some extent when studying differentiation processes. Second, gene regulation has multiple layers beyond direct TF-target regulation, including post-transcriptional regulations such as regulation by competing endogenous RNA, which may create confounding signals when ignoring them. Third, any GRN model is an abstraction of complex biophysical processes, and finding a good level of abstraction amenable to reverse engineering from experimental data certainly deserves further research. For example, ODE-based models like SCODE or GRISLI ignore non-linear effects and stochasticity, and it would be interesting to investigate whether more complex (or simpler) models could lead to computationally efficient procedures to reverse-engineer GRN more accurately. Fourth, TF-target regulation is known to be cell-type specific, which may penalize methods like GRISLI that infer a single GRN from data coming potentially from different cell types, and bias negatively the performance evaluation of GRN inference since our gold standard network contains regulations found in at least one cell type, which may not even be present in the expression dataset. Finally, GRN inference is essentially a problem of causal discovery, which is known to be fundamentally

more difficult than, for example, predicting the expression of a target genes from the expression of other genes; in particular, causal discovery may strongly benefit from interventional data involving TF knockdown as opposed to purely observational data.

A particular aspect of scRNA-seq data, compared to bulk transcriptomic data, is the so-called dropout effect which results in a possible inflation of zeros in the read count matrix (Kharchenko *et al.*, 2014). This may in turn penalize methods like GRISLI that do not model explicitly zero inflation, and may suggest extensions of the model to consider it. As a simple attempt to mitigate this problem, we tried to run GRISLI after imputing zero counts using ZINB-WaVE (Risso *et al.*, 2018) but found that the performance of GRN inference did not change or decreased after zero imputation (reaching 0.57 and 0.51, respectively, on the human and murine benchmarks). We leave it to future work to investigate further whether dropout deserves a specific treatment in GRN inference models.

As a last note, we wish to stress that GRISLI has two main parameters to tune ( $L$  and  $\alpha$ ), which have a significant impact on the performance. If some interactions are known, we suggest to tune them over a grid of candidate values by maximizing the fit between known and predicted interactions. Finding heuristics to automatically chose  $L$  and  $\alpha$  when no known interaction is available is an interesting future work.

## Acknowledgements

The authors thank deeply Héctor Climente-González and Samyadeep Basu for enlightening discussions, and two anonymous reviewers for useful comments and suggestions.

*Financial Support:* none declared.

*Conflict of Interest:* none declared.

## References

- Aibar, S. *et al.* (2017) Scenic: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
- Bach, F. *et al.* (2012) Structured sparsity through convex optimization. *Stat. Sci.*, **27**, 450–468.
- Bacher, R. and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.*, **17**, 63.
- Bastidas-Ponce, A. *et al.* (2019) Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, **146**, dev173849.
- Bergen, V. *et al.* (2019) Generalizing RNA velocity to transient cell states through dynamical modeling. 10.1101/820936.
- Bessonard, S. *et al.* (2014) Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development*, **141**, 3637–3648.
- Cannoodt, R. *et al.* (2016) Computational methods for trajectory inference from single-cell transcriptomics. *Eur. J. Immunol.*, **46**, 2496–2506.
- Chan, T.E. *et al.* (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.*, **5**, 251–267.e3.
- Chu, L.-F. *et al.* (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17**, 173.
- Deng, Q. *et al.* (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
- Filippi, S. and Holmes, C.C. (2017) A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Anal.*, **12**, 919–938.
- Haury, A.-C. *et al.* (2012) TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst. Biol.*, **6**, 145.
- Herbach, U. *et al.* (2017) Inferring gene regulatory networks from single-cell data: a mechanistic approach. *BMC Syst. Biol.*, **11**, 105.
- Hu, H. *et al.* (2018) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, **47**, D33–D38.
- Huynh-Thu, V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.
- Kanamori, M. *et al.* (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.
- Kharchenko, P.V. *et al.* (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.
- Kolodziejczyk, A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.
- La Manno, G. *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.
- Lönnberg, T. *et al.* (2017) Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci. Immunol.*, **2**, eaal2192.
- Maathuis, M.H. *et al.* (2010) Predicting causal effects in large-scale systems from observational data. *Nat. Methods*, **7**, 247–248.
- Marbach, D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.
- Matsumoto, H. *et al.* (2017) SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, **33**, 2314–2321.
- Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B*, **72**, 417–473.
- Moignard, V. *et al.* (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.*, **15**, 363–372.
- Moignard, V. *et al.* (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.*, **33**, 269–276.
- Neph, S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.
- Ocone, A. *et al.* (2015) Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, **31**, i89–i96.
- Patel, A.P. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
- Paul, F. *et al.* (2015) Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, **163**, 1663–1677.
- Risso, D. *et al.* (2018) A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.*, **9**, 284.
- Stegle, O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.
- Stergachis, A.B. *et al.* (2014) Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, **515**, 365–370.
- Tasic, B. *et al.* (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.*, **19**, 335–346.
- Trapnell, C. (2015) Defining cell types and states with single-cell genomics. *Genome Res.*, **25**, 1491–1498.
- Treutlein, B. *et al.* (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature*, **534**, 391–395.
- Zeisel, A. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.