STRUCTURE LEARNING FOR ZERO-INFLATED COUNTS WITH AN APPLICATION TO SINGLE-CELL RNA SEQUENCING DATA

BY THI KIM HUE NGUYEN[1,a], KOEN VAN DEN BERGE[2,c], MONICA CHIOGNA[3,d] AND DAVIDE RISSO[1,b]

[1]*Department of Statistical Sciences, Universiy of Padova,* [a]*nguyen@stat.unipd.it,* [b]*davide.risso@unipd.it*

[2]*Department of Applied Mathematics, Computer Science and Statistics, Ghent University,* [c]*koenvdberge@berkeley.edu*

[3]*Department of Statistical Sciences, University of Bologna,* [d]*monica.chiogna2@unibo.it*

The problem of estimating the structure of a graph from observed data is of growing interest in the context of high-throughput genomic data and single-cell RNA sequencing in particular. These, however, are challenging applications, since the data consist of high-dimensional counts with high variance and overabundance of zeros. Here we present a general framework for learning the structure of a graph from single-cell RNA-seq data, based on the zero-inflated negative binomial distribution. We demonstrate with simulations that our approach is able to retrieve the structure of a graph in a variety of settings, and we show the utility of the approach on real data.

1. Introduction. In recent years a growing interest has developed around the problem of retrieving, starting from observed data, the structure of graphs representing relationships among variables of interest. In fact, reconstruction of a graphical model, known as structure learning, traces back to the beginning of the 1990s, and a vast literature exists that considers the problem from various perspectives, within both frequentist and Bayesian approaches (see Drton and Maathuis (2017) for an extensive review). But a central role in the renewal of interest on structure learning has been played by molecular biology applications. In this field the abundance of data with increasingly large sample sizes, driven by novel high-throughput technologies, has opened the door for the development and application of structure learning methods, in particular when applied to the estimation of gene regulatory or gene association networks.

At the inception of transcriptomics, the technology of choice for measuring gene expression was the microarray assay, that, by optically scanning fluorophore intensities, provided data on a continuous scale (Irizarry et al. (2003)). When it came to (sparse) structure learning from these data, the first proposals assumed that data arose from a multivariate Gaussian distribution and took advantage of the many results and tools available for such family of distributions (see Junbai, Leo and Jan (2005), Peña (2008), Schäfer and Strimmer (2005), Yin and Li (2011), among others).

Later, a new technology allowed for the high-throughput sequencing of RNA molecules (i.e., RNA-seq) and quickly established itself as the reference technology for the study of genome-wide transcription levels (Wang, Gerstein and Snyder (2009)). One of the main advantages of RNA-seq over microarrays is that it allows to analyze small amounts of RNA, making it feasible to study gene expression even at single-cell resolution (Kolodziejczyk et al. (2015)). This new technology provided statisticians with a wealth of novel problems. Indeed, RNA-seq yields counts, rather than intensities on a continuous scale, as measures of gene expression. Data are usually high dimensional and, typically, come from skewed distributions with high variance. Moreover, they often show a large number of zeros, typically larger than expected under a Poisson or negative binomial model (Van De Wiel et al. (2013)).

Structure learning of graphs with such data was initially performed by exploiting data transformations, such as log, Box–Cox, copulas, etc (Abegaz and Wit (2015)). Although data transformation can work well in some circumstances, it can be also ill-suited, possibly leading to wrong inferences in some circumstances (Gallopin, Rau and Jaffrézic (2013)). Awareness of these problems fueled the development of methods for learning (sparse) graphical models tailored to count data. Allen and Liu (2013), Yang et al. (2013) and, more recently, Nguyen and Chiogna (2021) considered structure learning for Poisson and truncated Poisson counts. A general class of models was studied in Yang et al. (2015) which considered graphical models for the class of exponential family.

The challenges posed by RNA-seq technology are exacerbated in single-cell RNA sequencing (scRNA-seq). scRNA-seq allows the measurement of RNA from individual cells, promising to permit the study of gene interactions at an unprecedented resolution (McDavid et al. (2019)). Some scRNA-seq platforms employ *unique molecular identifiers* (UMIs) which help reduce amplification biases (Islam et al. (2014)) by counting unique RNA molecules rather than *reads* potentially representing the same molecule more than once. This implies that the distribution of the resulting data is substantially different: read-count data typically show larger counts than UMI data and a more pronounced bimodality (Svensson (2020)). Moreover, the small amount of RNA present in the cell and the technical limitations of the sequencing platforms (e.g., a limited number of sequenced reads per cell) lead to higher variance and larger fraction of zero counts compared to "bulk" RNA-seq (McDavid et al. (2019), Risso et al. (2018)). As a result, single-cell RNA-seq genewise data distributions are highly skewed and show an abundance of zero counts. Inference using Gaussian models is definitely infeasible, even after variance stabilizing transformations, and even models for count data may suffer from high false discovery rates (see Gallopin, Rau and Jaffrézic (2013) and Section 6). To account for zero inflation, McDavid et al. (2019) proposed a Hurdle model, equivalent to a finite mixture of singular Gaussian distributions. The authors' model, however, does not account for the count nature of the data.

From this quick tour on problems and methods, it appears evident that principled solutions to structure learning that account for the possibility of overdispersion and/or zero inflation are still lacking. In this paper we try to fill this gap. We present a general framework, based on the zero-inflated negative binomial distribution, for learning the structure of a graph from single-cell RNA-seq data. We focus in particular on UMI data, as its growing popularity suggests that the majority of future studies will employ this technology.

The remainder of this article is organized as follows. In Section 2 we introduce a motivating dataset; we describe our proposed model in Section 3 and our structure learning procedure in Section 4. One key question in the literature is whether zero inflation needs to be accounted for in the data; we offer our perspective in Section 5. After exploring the behavior of our method in simulated data in Section 6, we apply our proposal to real single-cell RNA-seq data in Section 7. Section 8 concludes the article with a discussion.

2. A motivating example: Single-cell gene expression in the olfactory epithelium.
Despite the distributional challenges described in the previous section, single-cell data offer an unprecedented opportunity to discover cellular dynamics, especially in developing cell populations. Graphical models could be an important tool to learn gene interactions from single-cell data, to learn how these change across conditions and throughout development and to identify potentially novel transcription factor target genes. While graphs are widely used in scRNA-seq to group similar cells in the space of gene expression, our approach learns a graphical model considering genes as nodes. This allows us to model cells as i.i.d. observations from a multivariate distribution in which the genes are the variables and the cells are considered a random sample from the cell population.

In particular, here we study gene expression from the mouse olfactory epithelium (OE). This tissue is made of two major mature cell types, olfactory neurons and sustentacular support cells. Furthermore, a stem cell niche provides a mechanism through which the tissue is regenerated (Fletcher et al. (2017), Gadye et al. (2017)). As the aim of the study is to understand how stem cells mature into neurons following tissue damage, we focused only on the cells in the neuronal lineage. Briefly, the olfactory reserve stem cells, called horizontal basal cells (HBCs), become activated and subsequently develop into globose basal cells (GBCs) and then into immature (iOSN) and finally mature olfactory neurons (mOSNs). By reconstructing the structure of the graph for each of these cell types separately, we hope to get a glimpse of the relationships between genes in neuronal development. For instance, the cell type that results in the most highly connected graph could indicate the most transcriptionally active developmental stage (Fletcher et al. (2017)).

3. Model specification.

3.1. *Preliminaries.* A probabilistic graphical model requires the definition of a pair, $(G, \mathcal{F})$ say. Here $G = (V, E)$ represents an undirected graph, where V is the set of nodes and $E = \{(s, t) : s, t \in V, s \neq t\}$ represents the set of undirected edges. Each node in the graph corresponds to a random variable $X_s, s \in V$; the existence of an edge $(s, t) \in E$ indicates the dependency of the random variables X_s and X_t. Moreover, $\mathcal{F}$ represents a family of probability measures for the random vector $\mathbf{X}_V$, indexed by V and with support $\mathcal{X}_V$.

Thanks to the well-known Markov properties (global, local, pairwise, see Lauritzen (1996)), the pattern of edges in the graph translates into conditional independence properties for variables in $\mathbf{X}_V$ which, in turn, allow possible factorizations of $\mathcal{F}$ into smaller, more tractable entities. In undirected graphical models, each absent edge (s, t) in E has the role of portraying the conditional independence,

$$X_s \perp\!\!\!\perp X_t | \mathbf{X}_{V \setminus \{s,t\}},$$

and the family $\mathcal{F}$ is said to satisfy the pairwise Markov property with respect to G. The smallest undirected graph G with respect to which $\mathcal{F}$ is pairwise Markov is given the name conditional independence graph.

When all variables in $\mathbf{X}_V$ are discrete with positive joint probabilities, as is the case of this paper, the three kinds of Markov properties are equivalent so that a factorization of the joint probability distribution with respect to the cliques (fully connected subsets of vertices) of the graph G is also guaranteed (Lauritzen (1996), Chapter 3).

3.2. *The model specification.* Let x_{is} be the gene expression for gene $s \in V$ in cell $i \in \{1, 2, \ldots, n\}$; we assume that the distribution of each variable X_{is}, conditional to all possible subsets of variables $\mathbf{X}_{iK}$, $K \subseteq V$ is a zero-inflated negative binomial (zinb) distribution,

$$(3.1) \quad \begin{aligned} &f_{\text{zinb}}(x_{is}; \mu_{is|K}, \theta_s, \pi_{is|K} | \mathbf{x}_{iK \setminus \{s\}}) \\ &= \pi_{is|K} \delta_0(x_{is}) + (1 - \pi_{is|K}) f_{nb}(x_{is}, \mu_{is|K}, \theta_s | \mathbf{x}_{iK \setminus \{s\}}), \end{aligned}$$

where $\delta_0(\cdot)$ is the Dirac function, $\pi_{is|K} \in [0, 1]$ is the probability that a 0 is sampled from a distribution degenerate at zero and $f_{nb}(\cdot, \mu, \theta)$ denotes the probability mass function of the negative binomial (NB) distribution with mean μ and inverse dispersion parameter θ. We assume that

$$(3.2) \quad \ln(\mu_{is|K}) = \nu^{\mu}_{s|K} + \sum_{t \in K \setminus \{s\}} \beta^{\mu}_{st|K} x_{it},$$

$$(3.3) \quad \text{logit}(1 - \pi_{is|K}) = \nu^{\pi}_{s|K} + \sum_{t \in K \setminus \{s\}} \beta^{\pi}_{st|K} x_{it}.$$

A missing edge between node s and node t corresponds to the condition $\beta^{\mu}_{st|K} = \beta^{\mu}_{ts|K} = \beta^{\pi}_{st|K} = \beta^{\pi}_{ts|K} = 0$, $\forall K \subseteq V \setminus \{s\}$. On the other hand, one edge between node s and node t implies that at least one of the four parameters $\beta^{\mu}_{st|K}, \beta^{\mu}_{ts|K}, \beta^{\pi}_{st|K}, \beta^{\pi}_{ts|K}$ is different from 0.

This specification defines a family of models that includes the most common models employed for count data and embraces a variety of situations. It is evident that, when $\pi_{s|K} = 0$, $\forall K \subseteq V \setminus \{s\}$, the model reduces to a NB distribution which, in turn reduces to a Poisson distribution when the inverse dispersion parameter θ_s tends to infinity. When $\pi_{s|K} > 0$, zero inflation comes into play, and zero-inflated Poisson and NB models can be considered. In this case, when $\beta^{\pi}_{st|K} = 0$, $\forall t \in K \setminus \{s\}$, the neighborhood of a node s is defined to be the set of effective predictors of $\mu_{s|K}$ and consists of all nodes t for which $\beta^{\mu}_{st|K} \neq 0$. On the other side, when $\beta^{\mu}_{st|K} = 0$, $\forall t \in K \setminus \{s\}$, the neighborhood of a node s is defined to be the set of effective predictors of $\pi_{s|K}$ and consists of all nodes t for which $\beta^{\pi}_{st|K} \neq 0$. In other words, the family includes models in which the structure of the graph is attributable only to one of the two parameter components, $\pi_{s|K}$ or $\mu_{s|K}$.

The difficulty with our model specification is that the definition of a set of conditional distributions does not guarantee the existence of a valid joint distribution, that is, a joint distribution that possesses the specified conditionals. This might create difficulties in interpreting the resulting graph in probabilistic terms: if the joint distribution does not exist, graphical separations stored in G, as a result of our model specification, might not correspond to conditional independence properties on $\mathcal{F}$. However, our formulation guarantees the existence of the joint distribution in a number of relevant subcases. In the following theorem, we clarify conditions for existence of a joint distribution coherent with the conditional specification (see Section 1.1, Supplementary Material (Nguyen et al. (2023)), for a proof).

THEOREM 1.   *Let* $\mathbf{X}_V = (X_1, X_2, \ldots, X_p)$ *be a p-random vector with support* $\mathcal{X}_V$. *Assume that a set of univariate conditional probability mass functions of the kind* (3.1) *are given for variables in* $\mathbf{X}_V$. *Then a joint distribution having those conditionals exists if and only if* θ_s *is constant for all* $s \in V$, *and all regression coefficients* $\beta^{\mu}_{st|K}$ *are negative,* $\forall K \subseteq V$.

The condition on negativity of local regression coefficients in (3.2) resembles a condition known in the literature of Markov random fields as "competitive relationship" (Besag (1974)). Generally speaking, the presence of only negative relations among entities is quite a rare event and incapability of capturing positive dependencies might be a severe drawback in various applications. Nevertheless, the existence of a joint distribution in these specific cases assures that statistical guarantees hold for conditional approaches to structure learning, such as the one used in this paper, and somehow softens the hazard of the use of such algorithms outside the conditions of existence of a joint distribution.

4. Structure learning.   A conditional independence graph $G = (V, E)$ on $\mathbf{X}_V$ can be estimated by estimating, for each node $s \in V$, its neighborhood. Hence, one can proceed by estimating the conditional distribution of $X_s | \mathbf{X}_{V \setminus s}$ and fixing the neighborhood of s to be the index set of variables $\mathbf{X}_{N(s)}$ on which the conditional distribution depends.

To estimate the neighborhood of each node, we employ the PC-stable algorithm, a variant of the PC algorithm first proposed by Spirtes, Glymour and Scheines (2000). The PC algorithm starts with a complete graph on V. Marginal independencies for all pairs of nodes are tested and edges removed when marginal independencies are found. Then, for every pair of linked nodes, independence is tested, conditional to all subsets of cardinality one of the adjacency sets of the two nodes. This testing procedure is iterated, increasing in turn the size of the conditioning sets, until this reaches its maximum limit or a limit imposed by the user. Reasons for choosing the PC algorithm are many, spanning from its consistency (assuming

no latent confounders) under i.i.d. sampling (Spirtes, Glymour and Scheines (2000)) to its ability to deal with a large number of variables and only moderately large sample sizes. The variant that we employ, PC-stable (Colombo and Maathuis (2014)), allows to control instabilities due to the order in which the conditional independence tests are performed. To perform the tests, deviance test statistics are employed for which a chi-squared asymptotic distribution can be obtained by standard asymptotic theory.

In what follows, let $\mathbb{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$ be the collection of n samples drawn from the random vectors $\mathbf{X}_V$, with $\mathbf{x}^{(i)} = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Starting from the complete graph, for each s and $t \in V \setminus \{s\}$ and for any set of variables $\mathbf{S} \subseteq \{1, \ldots, p\} \setminus \{s, t\}$, we test, at some prespecified significance level, the null hypothesis $H_0 : \beta_{st|K}^{\mu} = \beta_{ts|K}^{\mu} = \beta_{st|K}^{\pi} = \beta_{ts|K}^{\pi} = 0$, with $K = \mathbf{S} \cup \{s, t\}$. In other words, we test if the data support the existence of the conditional independence relation $X_s \perp\!\!\!\perp X_t | \mathbf{X_S}$. If the null hypothesis is rejected, there exists an edge (s, t) in the resulting graph. A control is operated on the cardinality of the set $\mathbf{S}$ of conditioning variables which is progressively increased from 0 to $p - 2$ or to m, $m < (p - 2)$.

Assume $X_s | \mathbf{x}_{K \setminus \{s\}} \sim \text{zinb}(X_s; \mu_{s|K}, \theta_s, \pi_{s|K} | \mathbf{x}_{K \setminus \{s\}})$, as in equation (3.1). The conditional log-likelihood for variable X_s, given $\mathbf{x}_{K \setminus \{s\}}$, is obtained by

$$(4.1) \qquad \ell_s(\mathbf{v}_{s|K}, \boldsymbol{\beta}_{s|K}, \theta_s) = \sum_{i=1}^{n} \ln f_{\text{zinb}}(x_{is}; \mu_{is|K}, \theta_s, \pi_{is|K} | \mathbf{x}_{K \setminus \{s\}}^{(i)}),$$

where $\mathbf{v}_{s|K}, \boldsymbol{\beta}_{s|K}$ are linked to $\boldsymbol{\pi}_{s|K}, \boldsymbol{\mu}_{s|K}$ through equations (3.2)–(3.3). The estimates $\hat{\mathbf{v}}_{s|K}$, $\hat{\boldsymbol{\beta}}_{s|K}, \hat{\theta}_s$ of the parameters $\mathbf{v}_{s|K}, \boldsymbol{\beta}_{s|K}, \theta_s$ are obtained by maximizing the conditional log-likelihood given in equation (4.1), that is,

$$(\hat{\mathbf{v}}_{s|K}, \hat{\boldsymbol{\beta}}_{s|K}, \hat{\theta}_s) = \underset{(\mathbf{v}_{s|K}, \boldsymbol{\beta}_{s|K}, \theta_s) \in \mathbb{R}^{2|K|+1}}{\operatorname{argmax}} \ell_s(\mathbf{v}_{s|K}, \boldsymbol{\beta}_{s|K}, \theta_s).$$

See Section 1.2, Supplementary Material (Nguyen et al. (2023)), for details on the estimation procedure. A deviance test statistic for the hypothesis $H_0 : \beta_{st|K}^{\mu} = \beta_{st|K}^{\pi} = 0$ can be obtained as

$$D_{s|K} = 2(\ell_s(\hat{\mathbf{v}}_{s|K}, \hat{\boldsymbol{\beta}}_{s|K}, \hat{\theta}_s) - \ell_s(\hat{\mathbf{v}}_{s|K}^0, \hat{\boldsymbol{\beta}}_{s|K}^0, \hat{\theta}_s^0)),$$

where $\hat{\mathbf{v}}_{s|K}^0, \hat{\boldsymbol{\beta}}_{s|K}^0, \hat{\theta}_s^0$ are the maximum likelihood estimates of the parameters under H_0. It is readily available that $D_{s|K}$ is asymptotically chi-squared distributed with two degrees of freedom under the null hypothesis, provided that some general regularity conditions hold:

REMARK 1. On assuming faithfulness of the node conditional distributions to the graph G, consistency of the algorithm can be proved in the case of competitive relationships in $\mu_s | K$ by suitably modifying results in Nguyen and Chiogna (2021). We recall that a distribution $P_{\mathbf{X}}$ is said to be faithful to the graph G if for all disjoint vertex sets $A, B, C \subset V$ it holds

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B | \mathbf{X}_C \quad \Rightarrow \quad A \perp\!\!\!\perp_G B | C,$$

where $A \perp\!\!\!\perp_G B | C$ means that A and B are separated in G by C. Thanks to the equivalence between local and global Markov properties, faithfulness of the local distributions guarantees faithfulness of the joint distribution.

REMARK 2. Although a theoretical proof of convergence of the algorithm is in question in the case of unrestricted relationships among variables, inference on the structure is still principled within a pseudo-likelihood perspective, that is, by approximating the likelihood function by a product of the conditional likelihood functions. Different pseudo-likelihood-based structure estimators have been shown to be consistent under a conditional model construction (see Csiszár and Talata (2006), among others); see also Nguyen and Chiogna (2021)

for an empirical exploration of convergence of a similar algorithm under the Poisson assumption in the case of unrestricted relationships among variables.

REMARK 3.   A large sample size, as typical in the applications at hand, impacts on the actual significance level of individual tests. Moreover, a multiplicity of tests are performed by the algorithm. For this reason we advise to set the nominal level of the test α to $\alpha_n = 2(1 - \Phi(n^b))$, where $0 < b < 1/2$ is related to the average neighborhood size. This choice is based on results in Nguyen and Chiogna (2021) and guarantees that the probability that a type I or II error occurs in the whole testing procedure goes to zero as $n \to \infty$; that is, it asymptotically controls the familywise error rate of all potential tests that could be done.

REMARK 4.   The chosen learning strategy has some advantages over alternative approaches based on sparse regressions (see also Nguyen and Chiogna (2021) for an extended discussion in the Poisson case). Sparsity can be easily implemented by a control on the conditional set size, instead of a control on parameter magnitudes, which can lead to overshrinkage. Moreover, it offers computational advantages, especially when sparse networks are the target of inference.

5. Zero inflation: A real issue?.   The need for modeling zero inflation in single cell data is a question at the core of an ongoing debate, with several authors arguing that the negative binomial distribution is sufficient to fit single-cell RNA-seq data when unique molecular identifiers are used (Sarkar and Stephens (2021), Svensson (2020), Townes et al. (2019), Vieth et al. (2017)). Indeed, the ability to distinguish between a nonzero-inflated distribution and zero-inflated alternatives highly depends on the relative size of the parameters of the distributions.

To gain a better understanding of this problem, we have tried to assess the misspecification cost due to assuming a zero-inflated distribution when no zero inflation occurs. To this aim, we confined ourselves to a univariate case with no covariates, fixed a nonzero-inflated model and measured the model misspecification cost occurring when using its zero-inflated counterpart by using the squared Hellinger distance as loss function. Such a loss function should, in principle, indicate, in an inferential sense, how far apart the two distributions are.

To this aim, let $\mathcal{Y} = \{0, 1, 2, \ldots, +\infty\}$ be the support of a discrete variable Y. We consider for Y a true probability distribution $P(y; \phi_0)$, $\phi_0 \in \Phi$ as well as a family $\mathcal{F} = \{Q(y; \psi), \ \psi \in \Psi\}$, $\Phi \subseteq \Psi$, of zero-inflated versions of the true probability distribution $P(y; \phi_0)$. The squared Hellinger distance between two probability distributions P and Q is defined as

$$d_h^2(P, Q) = \frac{1}{2} \sum_{y \in \mathcal{Y}} (\sqrt{p}_y - \sqrt{q}_y)^2$$

$$= 1 - \sum_{y \in \mathcal{Y}} \sqrt{p_y q_y},$$

where $p_y = P(y; \phi_0)$ and $q_y = Q(y; \psi)$.

In particular, assume that P is a NB distribution with $\phi_0 = (\mu_0, \theta_0)$ and Q is a zinb distribution, defined as

$$Q(y; \psi) = \pi \delta_0(0) + (1 - \pi) P(y; \phi_0),$$

with $\psi = (\phi_0, \pi)$. Hence, the squared Hellinger distance of P and Q can be written as

$$d_h^2(P, Q) = 1 - \sqrt{1 - \pi}(1 - P(0; \phi_0)) - \sqrt{P(0; \phi_0)}\sqrt{\pi + (1 - \pi) P(0; \phi_0)}.$$
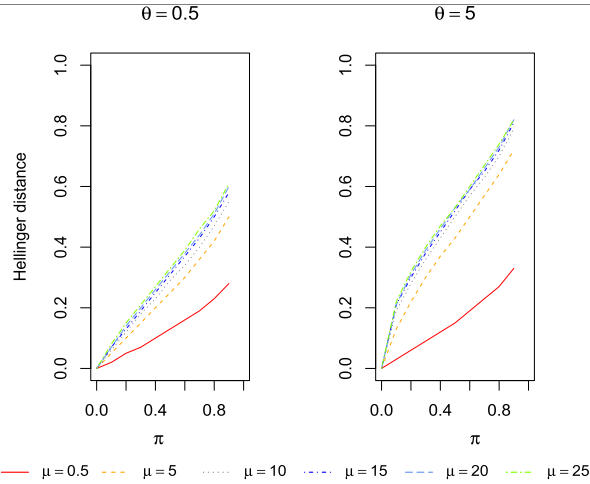
FIG. 1.   *Hellinger distance between zinb and NB distribution.*

Figure 1 and Supplementary Table S1 show the value of the Hellinger distance in a number of cases. As expected, the distance increases with the probability of zero inflation π. However, when the inverse dispersion parameter θ_0 and/or the mean μ_0 are small, the distance between the distributions is small, even in the case of moderate to large π. In fact, when μ_0 and θ_0 are both small (low mean and high variance), the two distributions are close, even when $\pi = 0.9$.

Broadly speaking, as maximum likelihood estimators and minimum Hellinger distance estimators are asymptotically equivalent, it emerges that, in inferential terms, the degree of zero inflation of a true model could be difficult to ascertain, as suitable choices of the parameters of the noncontaminated component may possibly absorb the excess of zeros generated by the contamination. This is despite identifiability of the zinb model (see Section 1.3, Supplementary Material (Nguyen et al. (2023)), for a proof). These remarks might contribute to the ongoing debate about existence of zero inflation from a novel perspective.

6. Simulations.   We devote this section to the empirical study of consistency of the proposed algorithms. In particular, we concentrate on the ability of proposed methods to recover the true structure of the graphs. We also list the running time of each algorithm. As measures of the test's accuracy, we adopt three criteria, including Precision P, Recall R and their harmonic mean, known as F_1-score, respectively, defined as

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = 2\frac{P.R}{P + R},$$

where TP (true positive), FP (false positive) and FN (false negative) refer to the number of inferred edges (Liu, Roeder and Wasserman (2010)).

The considered algorithms are listed below, along with specifications, if needed, of tuning parameters. For all PC-like algorithms, we let the maximum cardinality of conditional independence set be $m = 8$ for $p = 10$ and $m = 3$ for $p = 100$:

- *PC-zinb1*: zinb models in which the structure of the graph is attributable to both of the two parameter components $\mu_{s|K}$ and $\pi_{s|K}$;
- *PC-zinb0*: zinb models in which the structure of the graph is attributable to only the parameter component $\mu_{s|K}$ and consider $\pi_{s|K}$ as a constant (i.e., $\beta^\pi_{st|K} = 0$, $\forall t \in K \setminus \{s\}$, $\forall s \in V$);
- *PC-nb*: Negative binomial model, that is, the special case of zinb models where $\pi_{s|K} = 0$;
- *PC-pois*: Poisson model (Nguyen and Chiogna (2021)).

6.1. *Data generation.* For two different cardinalities ($p = 10$ and $p = 100$), we consider three graphs of different structure: (i) a scale-free graph in which the node degree distribution follows a power law, (ii) a hub graph, where each node is connected to one of the hub nodes and (iii) an Erdos–Renyi graph, where the presence of the edges is drawn from independent and identically distributed Bernoulli random variables.

To construct the scale-free and Erdos–Renyi networks, we employed the R package *igraph* (Csardi, Nepusz et al. (2006)). For the scale-free networks, we followed the Barabasi–Albert model with constant out-degree of the vertices $\nu = 2$ for $p = 10$ and $\nu = 0.2$ for $p = 100$. For the Erdos–Renyi networks, we followed the Erdos–Renyi model with probability to draw one edge between two vertices $\gamma = 0.3$ for $p = 10$ and $\gamma = 0.03$ for $p = 100$. To construct the hub networks, we assumed two hub nodes for $p = 10$ and five hub nodes for $p = 100$; see Supplementary Figure S2 and Supplementary Figure S3 for representative plots of the three chosen graphs for $p = 10$ and $p = 100$, respectively.

For the given graphs, 50 datasets were sampled with four different sample sizes, $n = \{100, 200, 500, 1000\}$ for $p = 10$, and three different sample sizes, $n = \{200, 500, 1000\}$ for $p = 100$. To generate the data, we followed the approach of the Poisson models in Allen and Liu (2013). Let $\mathbb{X} \in \mathbb{R}^{n \times p}$ be the set of n independent observations of random vector $\mathbf{X}$. Then $\mathbb{X}$ is obtained from the following model $\mathbb{X} = \mathbb{Y}A + \epsilon$, where $\mathbb{Y} = (y_{st})$ is an $n \times (p + p(p-1)/2)$ matrix whose entries y_{st} are realizations of independent random variables $Y_{st} \sim$ zinb(μ, θ, π) (or NB(μ, θ); or Pois(μ)) and $\epsilon = (e_{st})$ is an $n \times p$ matrix with entries e_{st} which are realizations of random variables $E_{st} \sim$ zinb($\mu_{\mathrm{nois}}, \theta, \pi$) (or nbinom($\mu_{\mathrm{nois}}, \theta$); or Pois($\mu_{\mathrm{nois}}$)). This approach leverages the additive property of these distributions and allows us to generate the required dependencies. In particular, let B be the adjacency matrix of a given true graph; then A takes the following form $A = [I_p; P \odot (1_p \operatorname{tri}(B)^T)]^T$. Here P is a $p \times (p(p-1)/2)$ pairwise permutation matrix, $\odot$ denotes the elementwise product and $\operatorname{tri}(B)$ is the $(p(p-1)/2) \times 1$ vectorized upper-triangular part of B (Allen and Liu (2013)).

6.2. *Results.* Figures 2 and 3 show the Monte Carlo means of the F_1-scores for each of the considered methods with $p = 100$ and low signal-to-noise ratio ($\mu_{\mathrm{noise}} = 0.5$), at high ($\mu = 5$) and low ($\mu = 0.5$) mean levels, respectively. Each value is computed as the average of the 50 values obtained by simulating 50 samples for the model corresponding to each network. Monte Carlo means of Precision P, Recall R and F_1-score are given in Supplementary Tables S2–S4.

The two values of $\mu = \{5, 0.5\}$ were chosen to mimic typical values observed in real full-length and droplet-based datasets, respectively. In fact, the mean expression level of transcription factors in the dataset presented in Section 2 is 0.67 (median 0.14), while the mean expression level of transcription factors in a similar experiment performed with a full-length protocol (Fletcher et al. (2017)) is 32.03 (median 7.89).

These results indicate that the PC-zinb1 algorithm and its variants (PC-zinb0, PC-nb, PC-pois) are consistent in terms of reconstructing the structure from given data. In fact, when the model is correctly specified, the F_1-scores of the algorithms are close to 1 when $n \geq 1000$ in all scenarios. This means that the proposed algorithm is able to recover the underlying graph from the given data for both low (Figure 3) and high (Figure 2) mean levels.

When the data are generated with a high mean level ($\mu = 5$), the PC-pois algorithm performs well only when it is the true model, that is, for data generated from Poisson random variables (Figure 2; Supplementary Table S4; Supplementary Figures S10 and S12). In the other scenarios, PC-pois often shows a low Precision (Figure 2; Supplementary Tables S2 and S3; Supplementary Figure S10). This result is expected since the node conditional Poisson distributions are unable to model the overdispersion generated by the (zero-inflated) negative binomial distributions.
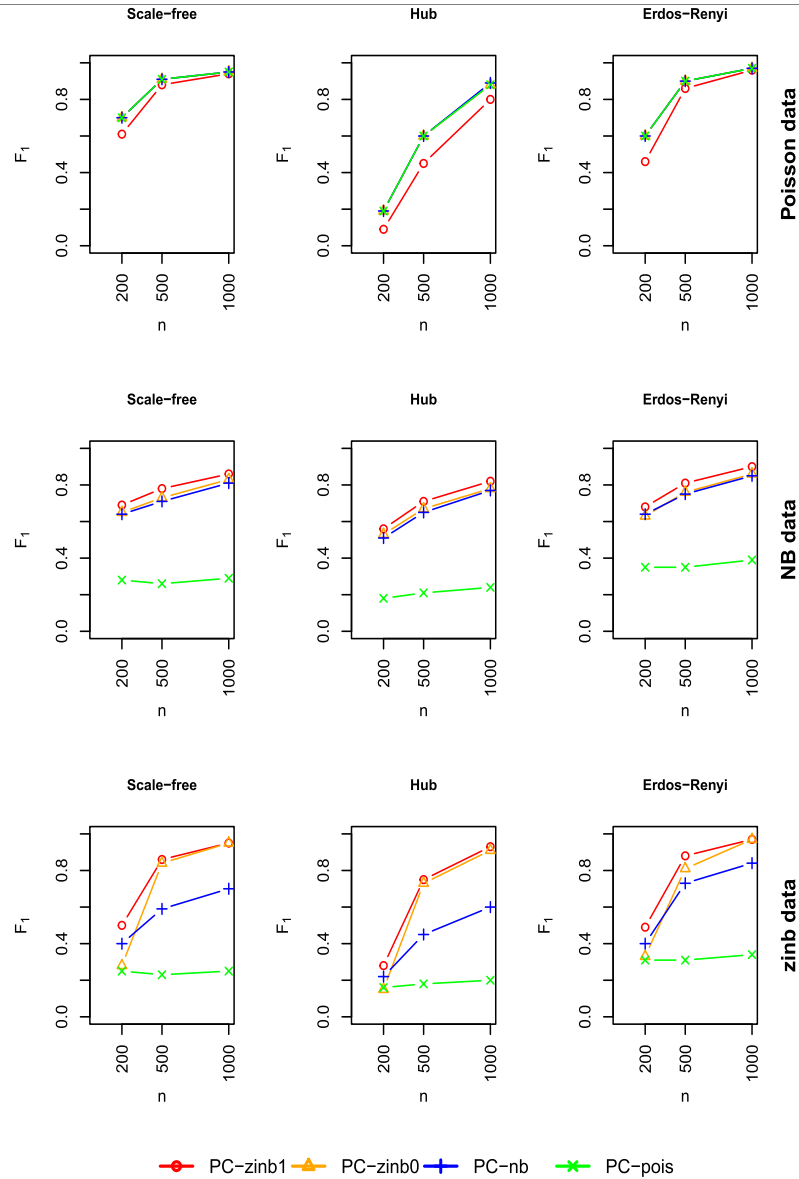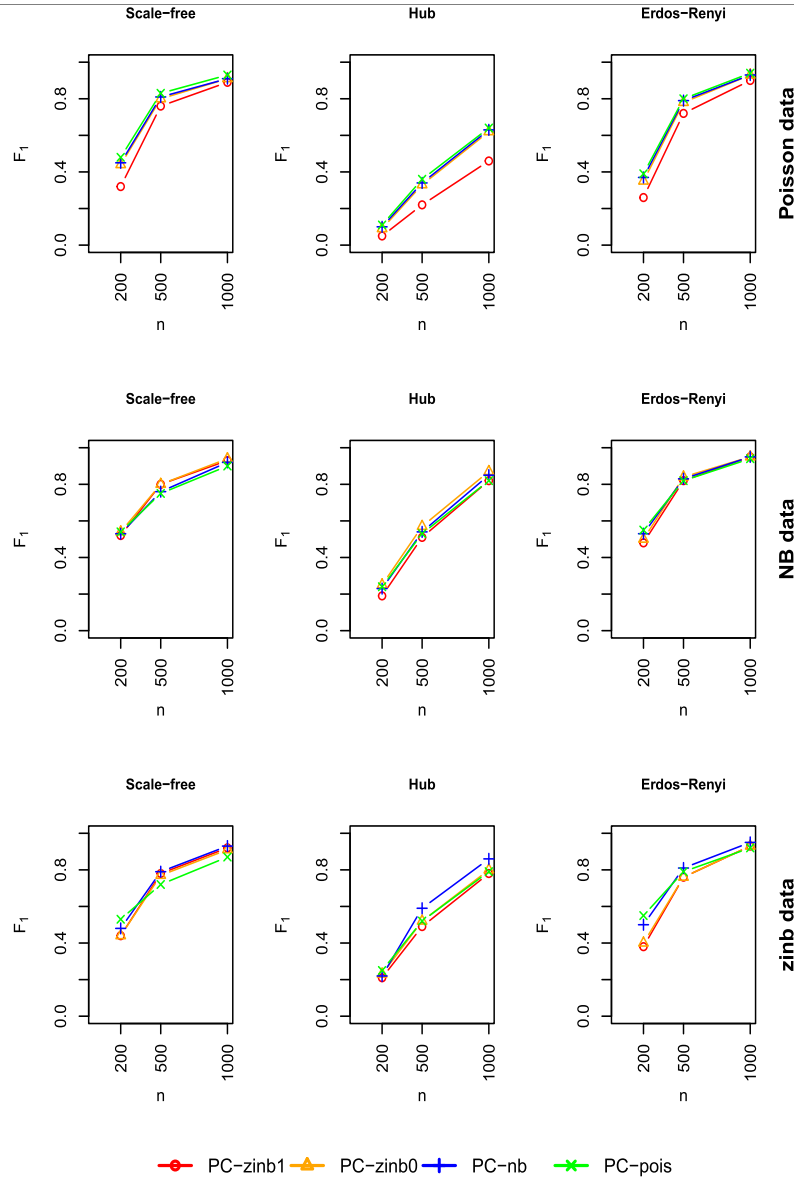
FIG. 2.   *F_1-score of the considered algorithms for the three types of graphs in Supplementary Figure S3 with $p = 100$, $\mu = 5$, $\theta = 0.5$, $\pi = 0.7$: scale-free; hub; Erdos–Renyi. The data were simulated from Poisson (top), NB (middle) and zinb (bottom) models. PC-zinb1: zinb model in which the structure of the graph is attributable to both of the two parameter components $\mu_{s|K}$ and $\pi_{s|K}$; PC-zinb0: zinb model in which the structure of the graph is attributable to only the parameter component $\mu_{s|K}$ and $\pi_{s|K}$ is constant; PC-nb: Negative binomial model, that is, the special case of zinb models where $\pi_{s|K} = 0$; PC-pois: Poisson model of Nguyen and Chiogna (2021).*

On the other end of the spectrum, the more general zinb models work well in all scenarios (Figure 2; Supplementary Tables S2–S4; Supplementary Figures S10 and S12). This is not surprising, as the data are generated according to models (e.g, Poisson, NB) that can be seen as special cases of the zinb distribution, which means that in all tested scenarios the zinb model is correctly specified.

The PC-nb algorithm, based on the negative binomial assumption, performs reasonably well (Figure 2; Supplementary Tables S2–S4; Supplementary Figures S10 and S12). However, in the hub graph (center column of Figure 2) its performances are slightly worse than the zinb models, showing low Precision when the true data generating distribution is node

FIG. 3. F_1-score of the considered algorithms for the three types of graphs in Supplementary Figure S3 with $p = 100$, $\mu = 0.5$, $\theta = 0.5$, $\pi = 0.7$: scale-free, hub, Erdos-Renyi. The data were simulated from Poisson (top), NB (middle) and zinb (bottom) models. PC-zinb1: zinb model in which the structure of the graph is attributable to both of the two parameter components $\mu_{s|K}$ and $\pi_{s|K}$; PC-zinb0: zinb model in which the structure of the graph is attributable to only the parameter component $\mu_{s|K}$ and $\pi_{s|K}$ is constant; PC-nb: Negative binomial model, that is, the special case of zinb models where $\pi_{s|K} = 0$; PC-pois: Poisson model of Nguyen and Chiogna (2021).

conditional zinb (Supplementary Figure S10; Supplementary Table S2). This result indicates that a zero-inflated negative binomial model may be needed when the mean is large (Risso et al. (2018)).

As we expected from the considerations reported in Section 5, the performances of the variants of PC-zinb are quite similar to each other when the mean and the dispersion parameter are both small, that is, when the data are characterized by low mean and high variance ($\mu = 0.5$, $\theta = 0.5$; Figure 3; Supplementary Tables S2–S4; Supplementary Figures S11 and S13). This might be explained by the fact that a suitable choice of the parameters may allow nonzero inflated models to absorb the excess of zeros (see Section 5 for more details).

Therefore, when applying our approach on this type of data, one should use the simplest variant, (i.e., PC-pois) to leverage the better computational performance (see last column of Supplementary Table S2–S4).

Moreover, we see no difference in the performance of the PC-zinb variants (PC-zinb1 and PC-zinb0). This is perhaps not surprising, as we simulated the same structure of the graph for both μ and π. These results suggest that the information inferred from μ is sufficient to reconstruct the correct graph in this case.

Finally, we compare the results to those obtained with the algorithm of McDavid et al. (2019) which employs a Gaussian Hurdle model (see Supplementary Table ??–??). The Hurdle model, applied to log transformed data shifted by 1, performs reasonably well only with a sufficient sample size ($n \geq 1000$) in the case of Erdos–Renyi and scale-free graphs but is unable to correctly reconstruct the hub graphs, even at large sample sizes. An extensive analysis of the results of the hub graph case revealed that the graph recovered by the Hurdle model is almost empty in a number of cases, especially at low sample sizes.

We have focused here on $p = 100$, as this setting is closer to our real application. The results for $p = 10$ are reported in Supplementary Figures S4–S9 and Supplementary Tables S5–S7 and lead to similar conclusions.

7. Results on real data. We demonstrate our method on the motivating example dataset described in Section 2. To this aim, we analyzed a set of cells, assayed with 10X Genomics (v2 chemistry) after injury of the OE, to characterize HBCs and their descendants during regeneration (Brann et al. (2020)). Starting from an initial set of 25,469 cells, low-quality samples as well as potential doublets were removed, as described in Brann et al. (2020). After clustering with the Leiden algorithm (Traag, Waltman and van Eck (2019)), known marker genes were used to identify cell types. We discarded the cell types outside of the neuronal lineage (macrophages, sustentacular cells and microvillar cells), obtaining a dataset consisting of 7782 HBCs, 5418 activated HBCs (HBC*), 755 GBCs, 2859 iOSN and 929 mOSN. For more details on the data preprocessing and cell annotation, see Brann et al. (2020).

We perform two complementary analyses on two different subsets of the dataset. First, we focus on transcription factor (TF) genes with the aim of identifying important networks of regulation in the different cell types that constitute the neuronal developmental lineage. We then turn our attention to the activated HBCs, a critical stage of neurogenesis, with the aim of identifying important transcription factors that regulate genes important for stem cell differentiation.

7.1. *Transcription factor genes.* Our first analysis focuses on the total set of 1543 known transcription factors in mouse which are thought to regulate the observed differentiation processes. Furthermore, we focus on the differentiation path, starting at the HBC* stage (i.e., activated stem cells upon injury) up to mature neurons, therefore, investigating the entire neuronal lineage in the trajectory of this dataset. As previously discussed in Section 2, we expect four different cell types along this path, being respectively HBC*, GBC, iOSN and mOSN, and we estimate the structure of the graph for each of these cell types. We selected the top 1000 cells with the highest means from the cell types that had more than 1000 cells (HBCs, HBC*, iOSN) to ensure a fair comparison between groups. In fact, the power of our algorithm to detect edges increases with the sample size, and since one of the goals of this analysis is to compare the graphs across cell types, we want to avoid a confounding effect due to the number of cells; see Supplementary Material (Nguyen et al. (2023)) Section 2 for details on the preprocessing.

The average degree of the graphs is highest at the activated stem cell stage, with an average degree of 4, and decreases as cells develop to mature neurons, with average degrees of 3.9,

3.3 and 3.5 for the GBC, iOSN and mOSN networks, respectively. To interpret the graph structure, we focus on the 2-core of each network, that is, we retain TFs that are associated with at least two other TFs, a preprocessing step that helps in understanding the core structure (Wang and Rohe (2016)).

We identify communities in each graph using the Leiden algorithm (Traag, Waltman and van Eck (2019)), and in order to validate the associations discovered by PC-zinb, we interpret each of the communities by computing overlaps with known functional gene sets in the MSigDB database (Subramanian et al. (2005), Liberzon et al. (2015)); see Supplementary Material (Nguyen et al. (2023)) Section 2 for details. The interpretation of these communities relies on known processes involved in the development of the olfactory epithelium, as found by previous research (e.g., Fletcher et al. (2017), Gadye et al. (2017)).

In the HBC* cell type, cells have been injured $\sim$ 24 h ago, so we expect response to injury and stem cells actively preparing for differentiation as well as replication to produce more stem cells to repair the epithelium. Four communities are discovered in the association network (Figure 4), broadly involved in either cell cycle, epigenetic mechanisms and (epithelial) cell differentiation (Supplementary Table S9). These communities reflect the need to divide in order to produce more cells, epigenetic mechanisms that are likely required to activate molecular processes upon injury and the differentiation of stem cells to restore the damaged epithelium.

In the GBC cell type, we expect cells to proliferate to produce immature neurons. We discover four communities (Figure 4), broadly involved in DNA replication, cell proliferation, signaling, expression regulation and cell differentiation (Supplementary Table S10). Relevant pathways, such as the P53 and notch signaling pathways, are also recovered for specific communities and have previously been found to be involved in neurogenesis in neuroepithelial stem cells (Marin Navarro et al. (2020), Wang et al. (2011)).

In the immature olfactory sensory neuron (iOSN) stage, we expect basal cells to start developing into immature neurons. Four communities are discovered (Figure 4), of which one community comprises the majority of the graph, that is, 63% of all TFs retained in the graph, and importantly is involved in neurogenesis (Supplementary Table S11). Other, smaller communities are enriched in processes such as cell and axon growth, wound healing, signaling and cell population maintenance.

Finally, in the mature olfactory sensory neuron stage (mOSN) we expect the final differentiation to functional neurons. Five communities are discovered (Figure 4), again with very different sizes. The largest communities are enriched in broader processes related to chromatin organization and transcription, possibly reflecting the basic changes required for cells to develop into and maintain at the mature stage (Supplementary Table S12). The third largest community is enriched specifically in the TGF-Beta pathway, known to be required for neurogenesis, and to modulate inflammatory responses (Meyers and Kessler (2017)).

Taken together, these results confirm previously known processes associated with differentiation of HBCs into mature neurons upon injury, with relevant processes highlighted by communities of transcription factors. Furthermore, while the community detection results are useful to validate the estimated graphs, they also provide a gateway to more detailed analysis, for example, investigation of hub genes (e.g., Chen et al. (2018)) or master regulators of development (e.g., Sikdar and Datta (2017)), therefore, unlocking powerful interpretation of single-cell RNA-seq datasets. We give an example of such detailed analysis in the next paragraph in which we focus on the role of the Trp63 TF in activated HBCs.

7.2. *Stem cell differentiation.*   Our second analysis focuses on a set of 242 genes, annotated with the term "stem cell differentiation" in the Mouse Genome Database (Bult et al.
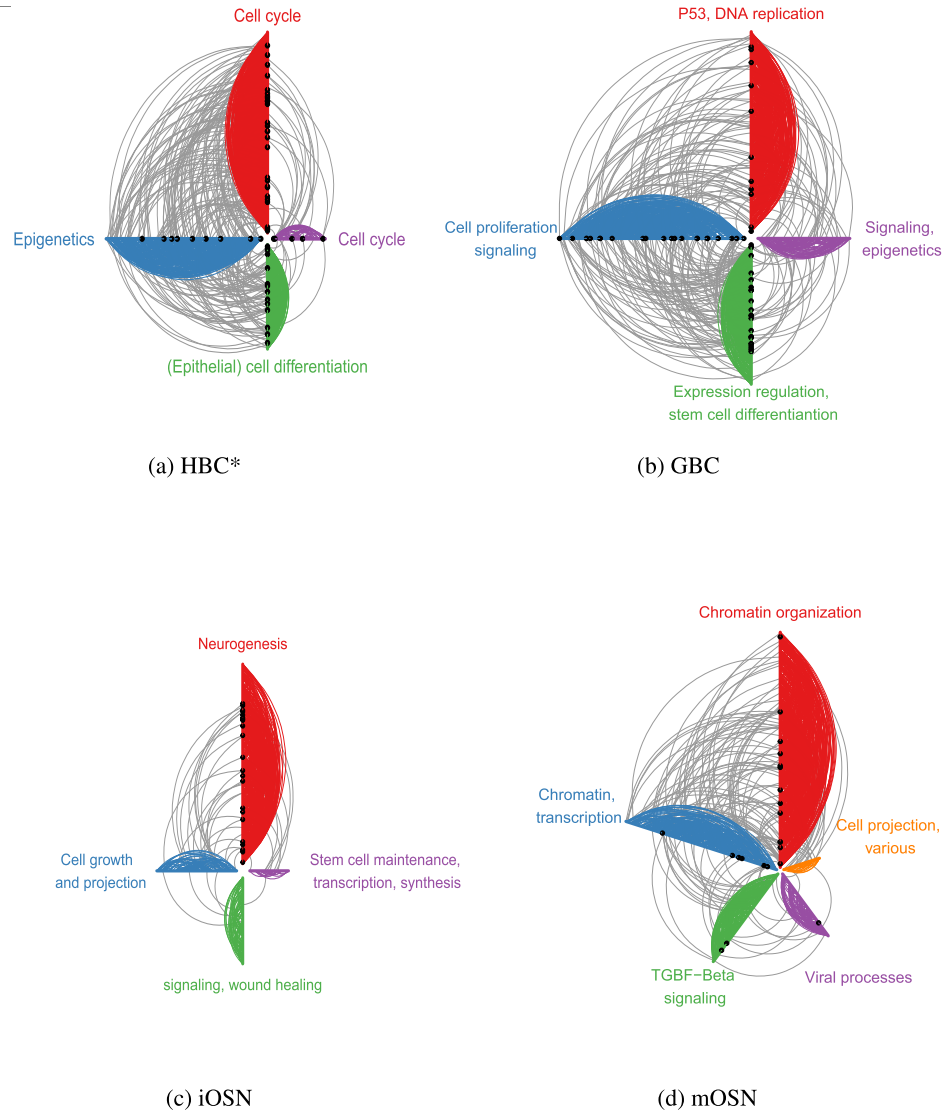
FIG. 4. *Hive plots (Bryan (2020), Krzywinski et al. (2012)) of TF gene networks estimated with PC-zinb. Gene communities were estimated using the Leiden algorithm and are represented on the axes of the plots and by different edge colors. The length of each axis is proportional to the size of the corresponding community; edges between two nodes in the same community are drawn in a community-specific color, while edges between two nodes in two different communities are colored in gray; hub nodes, defined as nodes with more than nine neighbors, are represented as solid black circles. Each axis (community) was annotated with the most enriched gene set (see Supplementary Material Section 2 (Nguyen et al. (2023))).*

(2019)), expressed in the activated HBC cell type. Following the same preprocessing employed for the first analysis and detailed in Section 2 of the Supplementary Material (Nguyen et al. (2023)), we obtain a dataset consisting of 1000 cells and 160 genes.

Our goal here is to infer the interactions among genes with a particular focus on the role of TFs in regulating target genes. Importantly, in this second analysis we include many genes that are not TFs, allowing us to focus on which genes are regulated by TFs at this specific point in development.

We expect to find several TFs as hub nodes in the graph. In fact, hub nodes, that is, nodes with a particularly high number of connections, may represent sites of signaling convergence, potentially indicating those genes that regulate other genes.
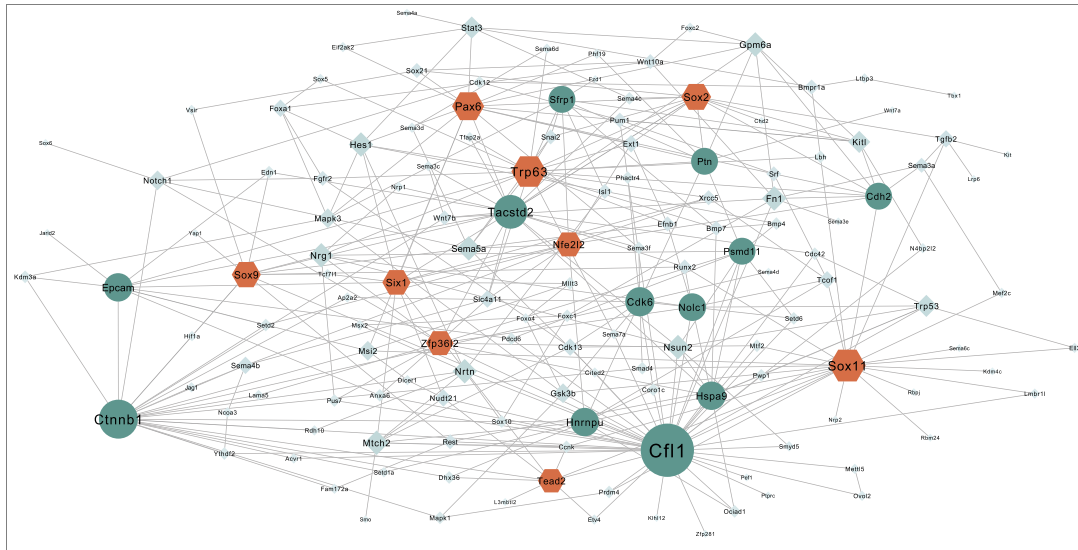
FIG. 5. *Network of stem cell differentiation gene set estimated with PC-zinb. Hub nodes are displayed with a circle, hub nodes that are also TF genes are displayed with a hexagon and the remaining nodes are displayed with a diamond.*

The PC-zinb algorithm inferred a sparse graph, shown in Figure 5, where hub nodes are displayed with a circle or a hexagon (when they are TFs). It is immediate to recognize important TFs previously demonstrated to be involved with stem cell differentiation, for example, *Trp63* (Senoo et al. (2007)), *Sox2* (Liu et al. (2013)) and *Sox9* (Jo et al. (2014)). Other hub nodes include genes that, while not TFs, have been shown to play a central role in this biological process. For instance, *Epcam* is known to be essential for the maintenance of selfrenewal in stem cells (González et al. (2009)). Another example is *Ptn*, the gene encoding the pleiotrophin growth factor, which has significant roles in cell growth and survival and has been demonstrated to be essential for stem cell maturation and neuronal development (Tang et al. (2019)).

We next focus on one of the most important TFs for stem cell activation, *Trp63*, by zooming in the subnetwork made of this gene and its direct neighbors (Figure 6). *Trp63* is one of the most important hubs in the network inferred by PC-zinb, with 20 connections. To validate the biological meaning of these connections, we leverage existing external data. In particular, Riege et al. (2020) performed a meta-analysis of 20 publicly available chromatin immunoprecipitation (ChIP-seq) datasets to create a curated catalog of p63 (the human ortholog of *Trp63*). Out of the 20 direct targets of *Trp63* in our network, 15 have been confirmed by
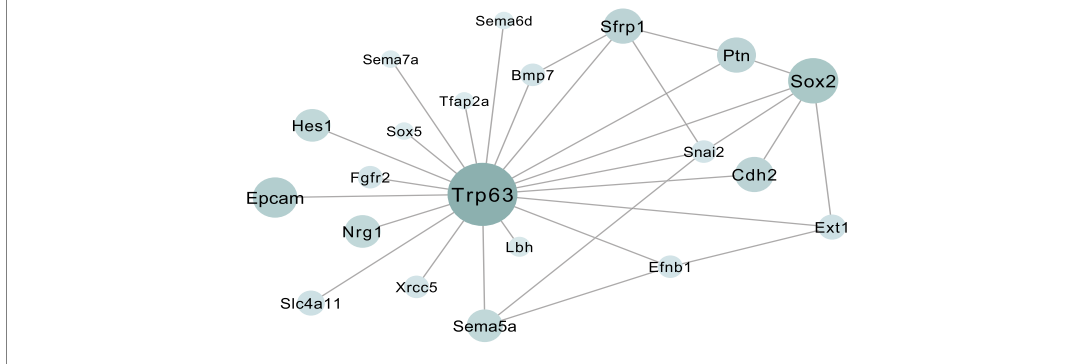

FIG. 6. *Subnetwork of Trp63 extracted from Figure 5.*

Riege et al. (2020) as direct targets of p63; that is, there is experimental evidence that the p63 protein binds either at the transcription start site (TSS) or upstream, indicating that p63 is either a promoter or enhancer of these genes (Supplementary File 3 of Riege et al. (2020)).

We want to stress that PC-zinb is able to find these putative TF-target pairs only on the basis of gene expression, hence proving itself as a useful tool to predict novel TF targets to be further validated with other techniques.

8. Discussion. In this work we have introduced PC-zinb, a class of constraint-based algorithms for structure learning, supporting possibly overdispersed and zero-inflated count data. In focusing on these two nonstandard but realistic situations, our framework goes beyond what has so far been proposed in the literature. Moreover, by leveraging the proposal in Nguyen and Chiogna (2021)—shown to be competitive with state-of-the-art methods supporting count data—we inherit the benefits of that approach, most notably, the existence of a theoretical proof of convergence of the algorithm under suitable assumptions, an easy implementation of sparsity by a control on the number of variables in the conditional sets and an invariance to feature scaling. On the synthetic datasets considered in Section 6, we showed that the algorithms work well in terms of reconstructing the structure from given data for large enough sample sizes, while providing biologically coherent information and insight on the real dataset analyzed in Section 7.

Our simulation studies allow us to derive various recommendations on the use of PC-zinb. Clearly, these do not rule out sensitivity analyses with respect to both model specification and tuning of the algorithms which remain an important part of the model criticism process. A control of the level of significance of the tests with respect to the sample size, n, and the expected size of the neighborhood of each node, b, is highly recommended to guarantee good reconstruction abilities. As in real applications, knowledge of the expected size of the neighborhood might be difficult to elicit; it may be prudent to try a range of values for b and check stability of results. This might also generate a sequence of models of decreasing complexity for increasing values of b and whose dynamic might also point researchers to the most significant connections.

If only the structure of the graph is of interest, irrespective of the strength of the links among variables, we suggest making use of the Poisson variant of the algorithm when the mean of the variables is small, so as to reduce computational complexity (Supplementary Tables S15 and S16). Moreover, when the mean of the variables is small, presence of zero inflation might not influence reconstruction abilities of the algorithms, as also confirmed by the small study on zero inflation in Section 5. In these situations we recommend using, at least in the first instance, nonzero-inflated models.

Clearly, in many applications learning the structure might not be the only goal, and one might want to gain a quantitative insight into the dependence structure of the underlying process by measuring the sign and the strength of the relations pictured in the graph. If the distribution needs also to be explicitly estimated, this can be achieved by using any of several existing parameter estimation methods conditional on the fixed structure learned by our approach.

If the null hypothesis $H_0 : \beta^{\mu}_{st|K} = \beta^{\pi}_{st|K} = 0$ fails to be rejected, PC-zinb will remove the edge between variables s and t. While such a procedure can only be justified in settings with high power, our simulation study shows that, even in settings with small sample sizes, our algorithm is able to achieve high power, and the correct underlying structure of the graph can be learned successfully.

While it is straightforward to interpret the case in which the neighborhood of s is defined by the predictors of $\mu_{s|K}$, that is, gene dependencies act on the average gene expression, the case of structure on $\pi_{s|K}$ requires more thought. If zero inflation represents true biological

signal, we can interpret a nonzero $\beta^{\pi}_{st|K}$ as the fact that the presence of gene t will influence the presence of gene s, regardless of their average expression. This is similar to McDavid et al. (2019). If zero inflation represents only technical noise, a simpler model with constant $\pi_{s|K}$ might be preferable. This is a special case of our general model. Since it is unclear what is the true nature of zero inflation in scRNA-seq data, we opted for generality in our model specification. Furthermore, having a general model expands the set of applications in which our approach may be useful.

The question of whether zero-inflated models are useful for the analysis of scRNA-seq data has been frequently posed in the recent literature. In Section 5 we try to shed some light on why a negative binomial distribution can fit UMI data well, as observed by Svensson (2020) and Sarkar and Stephens (2021) among others (see also our Figure 3). We show that, in the case of low mean and high variance, the zinb and NB distributions are very close to each other, rendering the question of whether UMI data are zero inflated or not. However, we also show that, in real data, zinb and NB models lead to different results, albeit with decent concordance between the inferred graphs (Supplementary Table S14). This result is only partially in agreement with those of Sarkar and Stephens (2021) in which the authors found that only a small percentage of genes show evidence of zero inflation. However, while Sarkar and Stephens (2021) focus much of their attention to the case of univariate gene expression, modeling zero inflation may be important when looking at correlation between genes (see also Yang and Ho (2022)).

Latent or unmeasured variables might induce associations between observed variables that can be spurious. Theoretical proposals are available to deal with the issue of latent factors in the setting in which the latent and observed variables are jointly Gaussian with the conditional statistics of the observed variables conditioned on the latent variables being specified by a graphical model (Chandrasekaran, Parrilo and Willsky (2012)), but to the best of our knowledge, no similar results are available for other families of models. For this reason in our paper, we simply leverage on convergence of the PC algorithm to the model marginalized over the latent factors.

As for the treatment of observed covariates and/or confounding factors, our proposed PC algorithm—that decomposes the structure learning problem into a series of tests performed on conditional log-likelihoods of each variable conditional on other variables—naturally allows to incorporate the covariates into the conditional regression models and, therefore, to estimate a covariate-adjusted structure for the graph. However, challenges remain if introduction of covariates augment the dimension of the conditional regression models to the point that one needs to resort to penalized tools. The treatment of both observed and latent covariates will be the object of future work.

Our real data analysis, aimed at assessing biological validity of the reconstructed network, has demonstrated the great importance of finding meaningful visualizations of large complex networks. Our proposal, based on a search for communities of variables and their association to gene ontologies via enrichment analysis, allowed us to confirm both biological interpretability of the estimated structure and to contribute to our understanding of which and where biological processes are occurring.

9. Software. The methods presented in this article are available in the *learn2count* R package, available at https://github.com/drisso/learn2count and as a zip file in the Supplementary Material (Nguyen et al. (2023)). The code to reproduce the analyses of this paper is available at https://github.com/drisso/structure_learning and as a zip file in the Supplementary Material (Nguyen et al. (2023)).

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/23-AOAS1732SUPPA; .pdf). Supplementary material includes proofs and mathematical details, details on the real data analysis, and supplementary figures and tables.

learn2count package (DOI: 10.1214/23-AOAS1732SUPPB; .zip). R package with the implementation of the methods proposed in this article.

Analysis code (DOI: 10.1214/23-AOAS1732SUPPC; .zip). Code to reproduce the analyses of this article.

REFERENCES

ABEGAZ, F. and WIT, E. (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Stat. Neerl.* **69** 419–441. MR3414705 https://doi.org/10.1111/stan.12066

ALLEN, G. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BRANN, D. H., TSUKAHARA, T., WEINREB, C., LIPOVSEK, M., VAN DEN BERGE, K., GONG, B., CHANCE, R., MACAULAY, I. C., CHOU, H. J. et al. (2020). Non-neuronal expression of Sars-CoV-2 entry genes in the olfaory system suggests mechanisms underlying Covid-19-associated anosmia. *Sci. Adv.* **6** eabc5801.

BRYAN, A. H. (2020). HiveR: 2D and 3D Hive Plots for R. R package version 0.3.63.

BULT, C. J., BLAKE, J. A., SMITH, C. L., KADIN, J. A., RICHARDSON, J. E. and THE MOUSE GENOME DATABASE GROUP (2019). Mouse genome database (MGD) 2019. *Nucleic Acids Res.* **47** D801–D806.

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

CHEN, D., YAN, W., FU, L. Y. and KAUFMANN, K. (2018). Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat. Commun.* **9** 4534. https://doi.org/10.1038/s41467-018-06772-3

COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

CSARDI, G., NEPUSZ, T. et al. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* **1695** 1–9.

CSISZÁR, I. and TALATA, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145. MR2275237 https://doi.org/10.1214/009053605000000912

DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393. https://doi.org/10.1146/annurev-statistics-060116-053803

FLETCHER, R. B., DAS, D., GADYE, L., STREET, K. N., BAUDHUIN, A., WAGNER, A., COLE, M. B., FLORES, Q., CHOI, Y. G. et al. (2017). Deconstructing olfactory stem cell trajectories at single-cell resolution. *Cell Stem Cell* **20** 817–830.

GADYE, L., DAS, D., SANCHEZ, M. A., STREET, K., BAUDHUIN, A., WAGNER, A., COLE, M. B., CHOI, Y. G., YOSEF, N. et al. (2017). Injury activates transient olfactory stem cell states with diverse lineage capacities. *Cell Stem Cell* **21** 775–790.

GALLOPIN, M., RAU, A. and JAFFRÉZIC, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS ONE* **8** e77503.

GONZÁLEZ, B., DENZEL, S., MACK, B., CONRAD, M. and GIRES, O. (2009). EpCAM is involved in maintenance of the murine embryonic stem cell phenotype. *Stem Cells* **27** 1782–1791. https://doi.org/10.1002/stem.97

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264. https://doi.org/10.1093/biostatistics/4.2.249

ISLAM, S., ZEISEL, A., JOOST, S., MANNO, G. L., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166. https://doi.org/10.1038/nmeth.2772

JO, A., DENDULURI, S., ZHANG, B., WANG, Z., YIN, L., YAN, Z., KANG, R., SHI, L. L., MOK, J. et al. (2014). The versatile functions of Sox9 in development, stem cells, and human diseases. *Genes Dis.* **1** 149–161.

JUNBAI, W., LEO, W. K. C. and JAN, D. (2005). New probabilistic graphical models for genetic regulatory networks studies. *J. Biomed. Inform.* **38** 443–455. https://doi.org/10.1016/j.jbi.2005.04.003

KOLODZIEJCZYK, A. A., KIM, J. K., SVENSSON, V., MARIONI, J. C. and TEICHMANN, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58** 610–620.

KRZYWINSKI, M., BIROL, I., JONES, S. J. and MARRA, M. A. (2012). Hive plots—rational approach to visualizing networks. *Brief. Bioinform.* **13** 627–644.

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York. MR1419991

LIBERZON, A., BIRGER, C., THORVALDSDÓTTIR, H., GHANDI, M., MESIROV, J. P. and TAMAYO, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* **1** 417–425. https://doi.org/10.1016/j.cels.2015.12.004

LIU, H., ROEDER, K. and WASSERMAN, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in Neural Information Processing Systems* 1432–1440.

LIU, K., LIN, B., ZHAO, M., YANG, X., CHEN, M., GAO, A., LIU, F., QUE, J. and LAN, X. (2013). The multiple roles for Sox2 in stem cell maintenance and tumorigenesis. *Cell. Signal.* **25** 1264–1271.

MARIN NAVARRO, A., PRONK, R. J., VAN DER GEEST, A. T., OLIYNYK, G., NORDGREN, A., ARSENIAN-HENRIKSSON, M., FALK, A. and WILHELM, M. (2020). P53 controls genomic stability and temporal differentiation of human neural stem cells and affects neural organization in human brain organoids. *Cell Death & Disease* **11** 52. https://doi.org/10.1038/s41419-019-2208-7

MCDAVID, A., GOTTARDO, R., SIMON, N. and DRTON, M. (2019). Graphical models for zero-inflated single cell gene expression. *Ann. Appl. Stat.* **13** 848–873. MR3963555 https://doi.org/10.1214/18-AOAS1213

MEYERS, E. A. and KESSLER, J. A. (2017). TGF-β family signaling in neural and neuronal differentiation, development, and function. *Cold Spring Harb. Perspect. Biol.* **9** a022244. https://doi.org/10.1101/cshperspect.a022244

NGUYEN, T. K. H. and CHIOGNA, M. (2021). Structure learning of undirected graphical models for count data. *J. Mach. Learn. Res.* **22** Paper No. 50, 53. MR4253743

NGUYEN, T. K, VAN DEN BERGE, K., CHIOGNA, M. and RISSO, D. (2023). Supplement to "Structure learning for zero-inflated counts with an application to single-cell RNA sequencing data." https://doi.org/10.1214/23-AOAS1732SUPPA, https://doi.org/10.1214/23-AOAS1732SUPPB, https://doi.org/10.1214/23-AOAS1732SUPPC

PEÑA, J. M. (2008). Learning Gaussian graphical models of gene networks with false discovery rate control. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. (J. H. Moore and E. Marchiori, ed.). *Lecture Notes in Computer Science* **4973**. Springer, Berlin.

RIEGE, K., KRETZMER, H., SAHM, A., MCDADE, S. S., HOFFMANN, S. and FISCHER, M. (2020). Dissecting the DNA binding landscape and gene regulatory network of p63 and p53. *eLife* **9**. https://doi.org/10.7554/eLife.63266

RISSO, D., PERRAUDEAU, F., GRIBKOVA, S., DUDOIT, S. and VERT, J. P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9** 1–17.

SARKAR, A. and STEPHENS, M. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53** 770–777. https://doi.org/10.1038/s41588-021-00873-4

SCHÄFER, J. and STRIMMER, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21** 754–764. https://doi.org/10.1093/bioinformatics/bti062

SENOO, M., PINTO, F., CRUM, C. P. and MCKEON, F. (2007). p63 is essential for the proliferative potential of stem cells in stratified epithelia. *Cell* **129** 523–536. https://doi.org/10.1016/j.cell.2007.02.045

SIKDAR, S. and DATTA, S. (2017). A novel statistical approach for identification of the master regulator transcription factor. *BMC Bioinform.* **18** 79. https://doi.org/10.1186/s12859-017-1499-x

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550. https://doi.org/10.1073/pnas.0506580102

SVENSSON, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38** 1–4.

TANG, C., WANG, M., WANG, P., WANG, L., WU, Q. and GUO, W. (2019). Neural stem cells behave as a functional niche for the maturation of newborn neurons through the secretion of PTN. *Neuron* **101** 32–44.

TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20** 1–16.

TRAAG, V. A., WALTMAN, L. and VAN ECK, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9** 5233. https://doi.org/10.1038/s41598-019-41695-z

VAN DE WIEL, M. A., LEDAY, G. G. R., PARDO, L., RUE, H., VAN DER VAART, A. W. and VAN WIERINGEN, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14** 113–128. https://doi.org/10.1093/biostatistics/kxs031

VIETH, B., ZIEGENHAIN, C., PAREKH, S., ENARD, W. and HELLMANN, I. (2017). powsimR: Power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33** 3486–3488. https://doi.org/10.1093/bioinformatics/btx435

WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. https://doi.org/10.1038/nrg2484

WANG, S. and ROHE, K. (2016). Discussion of "Coauthorship and citation networks for statisticians" [MR3592033]. *Ann. Appl. Stat.* **10** 1820–1826. MR3592035 https://doi.org/10.1214/16-AOAS977

WANG, W., LIU, W., WANG, Y., ZHOU, L., TANG, X. and LUO, H. (2011). Notch signaling regulates neuroepithelial stem cell maintenance and neuroblast formation in Drosophila optic lobe development. *Dev. Psychobiol.* **350** 414–428. https://doi.org/10.1016/J.YDBIO.2010.12.002

YANG, Z. and HO, Y.-Y. (2022). Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics* **78** 766–776. MR4450593 https://doi.org/10.1111/biom.13457

YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129 https://doi.org/10.1214/11-AOAS494