Joint gene network construction by single-cell RNA sequencing data

Meichen Dong1 | Yiping He2 | Yuchao Jiang1,3 | Fei Zou1,3

Abstract

In contrast to differential gene expression analysis at the single-gene level, gene regulatory network (GRN) analysis depicts complex transcriptomic interactions among genes for better understandings of underlying genetic architectures of human diseases and traits. Recent advances in single-cell RNA sequencing (scRNA-seq) allow constructing GRNs at a much finer resolution than bulk RNA-seq and microarray data. However, scRNA-seq data are inherently sparse, which hinders the direct application of the popular Gaussian graphical models (GGMs). Furthermore, most existing approaches for constructing GRNs with scRNA-seq data only consider gene networks under one condition. To better understand GRNs across different but related conditions at single-cell resolution, we propose to construct Joint Gene Networks with scRNA-seq data (JGNsc) under the GGMs framework. To facilitate the use of GGMs, JGNsc first proposes a hybrid imputation procedure that combines a Bayesian zero-inflated Poisson model with an iterative low-rank matrix completion step to efficiently impute zero-inflated counts resulted from technical artifacts. JGNsc then transforms the imputed data via a nonparanormal transformation, based on which joint GGMs are constructed. We demonstrate JGNsc and assess its performance using synthetic data. The application of JGNsc on two cancer clinical studies of medulloblastoma and glioblastoma gains novel insights in addition to confirming well-known biological results.

KEYWORDS
imputation, joint gene network, single-cell RNA sequencing, sparsity

1 | INTRODUCTION

Deciphering the landscape of transcriptomic interactions among genes (Trapnell et al., 2014) is essential in understanding the mechanism of complex diseases (Buil et al., 2015). Gene regulatory network (GRN) analysis is an attractive way to build transcriptional relationships among genes, providing a comprehensive view of gene expression dependencies (Langfelder and Horvath, 2008; Van de Sande et al., 2020). In a GRN, individual genes are represented as nodes, and gene pairs with a coexpression relationship are connected by edges (Chen and Mar, 2018). One focus area in biological and clinical research is the study of GRN changes across different tissues, cell types/states, and conditions. Compared to estimating a single network that applies to multiple conditions, separately constructing graphical models under different conditions is more robust and less biased (Lee et al., 2018), but its performance may be improved by utilizing the network similarity across the related conditions. As such,
GGMs and their adaptive models are popular for constructing GRNs using microarray data due to their ease of interpretation. Recently, bulk RNA-seq data and the newly emerging scRNA-seq data enable researchers to explore the gene transcriptional relationships at a finer resolution with a higher throughput; this enhances our understanding of transcriptional mechanisms and allows discoveries of novel therapeutic targets. Many network construction methods (Karlebach and Shamir, 2008; Marbach et al., 2012) have been developed for bulk RNA-seq data, including the popular GGMs. However, they cannot be readily applied to scRNA-seq data (Chen and Mar, 2018). Though both are count data, scRNA-seq data are inherently sparser, with an excessive amount of zeros due to either technical artifacts (false zeros), known as “dropouts” (Hicks et al., 2018), or biological absence of expression (true zeros) (Jiang et al., 2017; Dong and Jiang, 2019). Hence, traditional data processing and normalization methods (e.g., log-transformation) that work well for bulk RNA-seq data no longer work for the scRNA-seq data (Townes et al., 2019). Furthermore, most existing GRN models for scRNA-seq data are developed for a homogeneous cell population under one condition (Blencowe et al., 2019; Pratapa et al., 2020), with the exception of Wu et al. (2020), who propose a Gaussian copula graph model for joint GRN construction using scRNA-seq data. However, the model directly applies the nonparanormal transformation (Liu et al., 2009) to the count data, which, as mentioned above, may not be appropriate for the single-cell setting. To properly employ the Gaussian copula model, for bulk RNA-seq data, Jia et al. (2017) propose to impute and continuize the read counts using a Poisson hierarchical model.

Inspired by Jia et al. (2017), we propose to impute and continuize the scRNA-seq count data by a Bayesian zero-inflated Poisson (ZIP) model in which the zero-inflation part accommodates and predicts the dropout events. The choice of the Poisson distribution is supported by the increasing evidence of empirical studies (Townes et al., 2019; Kim et al., 2020; Svensson, 2020). After the dropout zeros are initially imputed with the Bayesian ZIP model, we further improve the imputation with McImpute (Mongia et al., 2019) to borrow information across genes in realizing that the gene-specific ZIP model does not utilize information across genes. McImpute imputes the dropout events based on low-rank matrix completion, yet it may oversimplify the imputation procedure as it does not employ the gene expression distribution information. Upon completion of the within- and between-gene imputation procedures, we proceed to construct the joint GGMs.

Together, we introduce a comprehensive tool for constructing Joint Gene Networks with scRNA-seq (JGNSc) data from different but related conditions. The framework consists of three major steps as shown in Figure 1: (1) a hybrid iterative imputation and continuization procedure; (2) a nonparanormal transformation of the processed data from step (1); and (3) joint GRNs construction that employs the fused lasso technique (Danaher et al., 2014). There are two tuning parameters associated with the model in step (3). Hence, we also investigate the performances of several tuning parameter selection criteria, including Akaike information criterion (AIC) (Akaike, 1998), Bayes information criterion (BIC) (Schwarz et al., 1978), extended Bayesian Information criterion (EBIC) (Chen and Chen, 2008), and the stability approach to regularization selection (StARS) (Liu et al., 2010).

This research is motivated by a medulloblastoma scRNA-seq dataset from Hovestadt et al. (2019) and a glioblastoma (GBM) dataset from Neftel et al. (2019), where single-cell samples from multiple tumor subtypes are available. The research interest is to learn the transcriptional relationships among genes across different tumor subtypes, and thus to guide hypothesis generation for clinical/biological experiments. The rest of the paper is organized as follows: we present the proposed method in Section 2, show results from empirical evaluations and real data analysis in Section 3, and conclude in Section 4.

2 | METHODS

2.1 | Hybrid imputation procedure for scRNA-seq data

The proposed Bayesian framework largely follows Jia et al. (2017); we replace the Poisson model with the ZIP model to account for the dropout events in the scRNA-seq data. For gene $i$ ($1 \leq i \leq G$) and cell $j$ ($1 \leq j \leq n$), let the observed expression read count be $\{y_{ij}\}$. We define a latent variable $z_{ij}$ to indicate whether the read count is from a “dropout” event (i.e., zeros due to technical artifacts) or from a Poisson distribution (i.e., zeros and non-zeros from Poisson sampling):

$$z_{ij} = \begin{cases} 0 & \text{if } y_{ij} \text{ is from a dropout event due to technical artifacts,} \\ 1 & \text{if } y_{ij} \text{ is from a Poisson distribution.} \end{cases}$$

The library size of cell $j$ is $\tau_j = \sum_{i=1}^{G} y_{ij} / \text{median}(\sum_{i=1}^{G} y_{ij})$. Assuming the gene-specific
non-dropout rate, \( p_i \), is constant across cells from a homogeneous group, the ZIP model is

\[
z_{ij} | p_i \sim \text{Bernoulli}(p_i),
\]

\[
y_{ij} | \theta_i, z_{ij}, \tau_j \sim \begin{cases} \delta(0), & \text{if } z_{ij} = 0 \\ \text{Poisson}(\theta_i \tau_j), & \text{if } z_{ij} = 1, \end{cases}
\]

where \( \delta(0) \) is the Dirac delta function with a unit mass concentrated at zero and \( \theta_i \) is the library-size-adjusted baseline expression for gene \( i \). The priors for \( \theta_i \) and \( p_i \) are

\[
p_i \sim \text{Beta}(a_1, b_1), \quad \theta_i \sim \text{Gamma}(\alpha_i, \beta_i).
\]

\[
\alpha_i \sim \text{Gamma}(a_2, b_2), \quad \beta_i \sim \text{Gamma}(a_3, b_3).
\]

That is, the prior for the Bernoulli probability \( p_i \) follows a Beta distribution with hyperparameters \( \{a_1, b_1\} \), and the prior for the mean gene expression \( \theta_i \) follows a Gamma distribution with two gene-specific parameters \( \alpha_i \) and \( \beta_i \). The priors for \( \alpha_i \) and \( \beta_i \) are further assumed to be Gamma distributions and are a priori independent with hyperparameters \( \{a_2, b_2\} \) and \( \{a_3, b_3\} \), respectively. The full conditional posterior distributions for the parameters of interest are given below (see Web Supporting Information for more details):

\[
\theta_i | p_i, \alpha_i, \beta_i, y_{ij}, z_{ij}, \tau_j \sim \text{Gamma}\left(\alpha_i + \sum_{j=1}^{n} y_{ij} z_{ij}, \beta_i + \sum_{j=1}^{n} \tau_j z_{ij}\right).
\]

\[
z_{ij} | \theta_i, p_i, \alpha_i, \beta_i, y_{ij}, \tau_j \sim \text{Bernoulli}\left(\frac{p_i e^{-\theta_i}}{p_i e^{-\theta_i} + (1 - p_i) I(y_{ij} = 0)}\right).
\]

For parameter estimation and inference, we employ the Metropolis–Hastings-based Markov chain Monte Carlo (MCMC) algorithm as follows. Let \( t = 1, ..., T \) index the iterations in the MCMC procedure. Following Lemma 1 and Lemma 2 from Jia et al. (2017), we set \( a_2, a_3 \) to be small positive values, and \( b_2, b_3 \) to be large positive values, and increase \( b_2, b_3 \) in iteration \( t \) by:

\[
b_{2(i)} = b_{2(i-1)} + \frac{1}{t}, \quad b_{3(i)} = b_{3(i-1)} + \frac{1}{t}.
\]

We consider \( y_{ij} \) as a dropout event if the average value \( \sum_{i=1}^{T} z_{ij} / T \) is smaller than a prechosen threshold (0.75 by default), and impute it to be \( \hat{\theta}_i \times \tau_j \), where \( \hat{\theta}_i \) is the posterior mean or mode. Note that instead of imputing all zero events as done in Jia et al. (2017) for bulk RNA-seq data, we choose to only impute the zeros that are predicted to be dropouts, while retaining the true biological zeros. We let the imputed matrix from the posterior sampling via MCMC be \( \hat{\Theta} \).

In realizing that \( \hat{\theta}_i \) is estimated one gene at a time, we introduce a hybrid imputation procedure that integrates the Bayesian ZIP model and McImpute (Monga et al., 2019) to further improve the imputation accuracy. McImpute is a low-rank matrix completion method for imputing scRNA-seq data. It borrows expression information across genes and cells while performing the imputation with the following objective function:

\[
\min_{M^*} \text{rank}(M^*), \quad \text{such that } M = A(M^*),
\]

where \( A(\cdot) \) is a binary mask function indicating whether the element in matrix \( M \) is zero or not, and \( M^* \) is the complete expression matrix. As suggested by Mongia et al. (2019), this problem can be further converted into a nuclear norm minimization problem and be solved by invoking the Majorization-Minimization algorithm (Sun et al., 2016). We reimpute the dropout values in \( \hat{\Theta} \) iteratively by McImpute (Monga et al., 2019), as described in Algorithm 1. Within each iteration, \( \alpha \% \) of MCMC-imputed dropout values are masked back to zeros. The impact of \( \alpha \) is investigated via simulations.
ALGORITHM 1 JGNsc Hybrid imputation algorithm

1: **Input:** Single-cell gene expression matrix $Y$, number of iterations $B$.
2: **Output:** Imputed matrix $\Theta^*$.
3: **procedure** ITERATIVE IMPUTATION
   • Step 1: Run Bayesian ZIP model and obtain the imputed matrix as $\hat{\Theta}$.
   • Step 2: Run iterative McImpute.
4:   **for** $b = 1, \ldots, B$ **do**
5:     **Mask** $\alpha\%$ of the MCMC-imputed values in $\hat{\Theta}$ back to zero, and define
6:     the new data matrix as $M$.
7:     Perform McImpute on $M$ and derive the imputed matrix $\Theta^*_b$.
8:   **end for**
9: **Calculate** $\Theta^* = \frac{1}{B} \sum_{b=1}^B \Theta^*_b$.
10: **Return** $\Theta^*$
11: **end procedure**

### 2.2 Joint graphical lasso model

Prior to building the joint GGMs, we apply the nonparanormal transformation (Liu et al., 2009) to the final imputed data. The nonparanormal transformation is a popular procedure for preprocessing non-Gaussian distributed data that follow a nonparanormal distribution (Jia et al., 2017; Wu et al., 2020). Specifically, a random vector $X = (X_1, \ldots, X_G)^T$ follows a nonparanormal distribution $X \sim NPN(\mu, \Sigma, f)$, if there exist functions $\{f_j : j = 1, \ldots, G\}$ such that $Z \equiv f(X) \sim N(\mu, \Sigma)$, where $f(X) = \{f_1(X_1), \ldots, f_G(X_G)\}$. When $\{f_j\}$ is monotonic and differentiable, the nonparanormal distribution $NPN(\mu, \Sigma, f)$ is a Gaussian copula, and the conditional independence structure of the original graph is preserved in the precision matrix $\Omega = \Sigma^{-1}$.

After the imputation and the Gaussian transformation, we then employ the joint graphical lasso (JGL) (Danaher et al., 2014), which extends GGMs to simultaneously estimate multiple related precision matrices. For $K$ groups of Gaussian data, denote the precision matrix under condition $k (1 \leq k \leq K)$ as $\Omega^{(k)}$, with $\omega^{(k)}_{uv}$ being its $(u,v)$th entry. The objective function of JGL is:

$$
\max_{\Omega^{(k)}} \left[ \sum_{k=1}^K n_k \left\{ \log( |\Omega^{(k)}| ) - \text{trace}(S^{(k)} \Omega^{(k)}) \right\} - P(\Omega^{(k)}) \right],
$$

where $n_k$ is the number of observations under condition $k$, $S^{(k)} = (Y^{(k)})^T Y^{(k)}/n_k$ is the empirical covariance matrix for the $k$th expression dataset, and $P(\Omega^{(k)})$ is a convex penalty function that is defined as

$$
P(\Omega^{(k)}) = \lambda_1 \sum_{k=1}^K \sum_{u \neq v} |\omega^{(k)}_{uv}| + \lambda_2 \sum_{k<k'} \sum_{u \neq v} |\omega^{(k)}_{uv} - \omega^{(k')}_{uv}|$$

under the fused lasso framework. The penalty function involves two tuning parameters $\lambda_1$ and $\lambda_2$: $\lambda_1$ ensures a sparsity solution of the graphical lasso model; $\lambda_2$ enforces the parameters to be shared across conditions.

For tuning-parameter selection, many different criteria are available. In practice, the choice of a criterion is largely driven by data and research goals (Wysocki and Rhemtulla, 2019), and there is currently no gold standard for scRNA-seq data. Via simulations, we evaluate four popular model selection criteria that are adapted to the JGL models, including AIC, BIC, EBIC, and StARS with detailed definitions given below:

$$
\text{AIC}(\lambda_1, \lambda_2) = \sum_{k=1}^K \left[ n_k \text{tr}(S^{(k)} \hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) - n_k \log\det(\hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) \right] + 2E_k,
$$

$$
\text{BIC}(\lambda_1, \lambda_2) = \sum_{k=1}^K \left[ n_k \text{tr}(S^{(k)} \hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) - n_k \log\det(\hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) \right] + \log(n_k)E_k,
$$

$$
\text{EBIC}(\lambda_1, \lambda_2) = \sum_{k=1}^K \left[ n_k \text{tr}(S^{(k)} \hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) - n_k \log\det(\hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}) \right] + (\log(n_k) + 2\gamma\log(G)) \times E_k,
$$

where $E_k$ is the number of nonzero elements in $\hat{\Omega}^{(k)}_{\lambda_1, \lambda_2}$, $0 \leq \gamma \leq 1$ is a parameter for EBIC and is set to 0.5 in simulation to control false discovery rate while maintaining positive selection rate. We adapt the StARS method to the JGL models by optimizing one parameter at a time iteratively with details outlined in the Web Supporting Information.

### 2.3 Simulation of scRNA-seq data under different conditions

To simulate scRNA-seq data under different but related conditions, we first simulate conditionally independent structures based on the inverse nonparanormal transformation algorithm developed by Yahav and Shmueli (2012), which is later applied by Jia et al. (2017) according to the
simulation scheme below. For each condition $k$ ($1 \leq k \leq K$),

(a) Randomly sample from a multivariate Gaussian distribution with a known precision matrix $\Omega^{(k)}$ (details available in the Web Supporting Information). Denote the random samples by $X_1, \ldots, X_G$, where each variable $X_i = (X_{i1}, \ldots, X_{in})^T$ consists of $n$ realizations.

(b) For each variable $X_i$, derive the empirical CDF based on the $n$ realizations, and then calculate the cumulative probability for each $X_{ij}$.

(c) Generate the $k$th scRNA-seq count variables with prespecified parameters $X_1^* : \ldots, X_G^*$, where $X_i^* = (X_{i1}^*, \ldots, X_{in}^*)^T$ consists of $n$ realizations. $X_{ij}^*$ is sampled from a Poisson distribution with true mean level $\theta_i$, which is then assumed to be generated from a Gamma distribution. Derive the empirical CDF for each $X_i^*$, and then calculate the cumulative probability for each $X_{ij}^*$.

(d) Map the quantiles of the data points in $X_i^*$, which is generated from (c), to the cumulative probabilities calculated from (b). Denote the mapped counts for variable $i$ as $Y_{ij}$.

(e) Based on the zero-inflation parameter distribution, we simulate the random dropout event for each data point $z_{ij}$. The final count matrix will be presented as $\{z_{ij} \times Y_{ij}\}$.

3 | RESULTS

3.1 | Simulation settings and results

We simulate scRNA-seq data under $K = 2$ conditions following the strategy described in Section 2.3. The number of genes is set to $G = 100$. We vary the sample size, the precision matrix structure, and the dropout rate. As $n < G$ and $n \geq G$ are both possible in real single-cell studies, we set $n_1 = n_2 = n \in \{50, 100, 200, 300, 500\}$ for the two related conditions. To evaluate the impact of the mask rate $\alpha$ on the performance of JGNsc, we also vary $\alpha$ at $\{5\%, 10\%, \ldots, 95\%\}$. Web Figure 1 shows that JGNsc performs the best when $\alpha$ falls between 10% and 20%. The reported analysis results from here on are based on the analysis with an $\alpha$ of 15%. We investigate the effect of the precision matrix structures in three scenarios: (i) partially identical precision matrix structures and weights; (ii) identical precision matrix structure with different weights; and (iii) different precision matrix structures and weights. Here, the weights refer to the partial correlations between gene pairs, and the structures refer to the connections between gene pairs. In greater details, for scenario (i), we simulate two cases where the first 20 (case 1) or 50 (case 2) genes under the two conditions are set to have different weights and precision matrix structures, but the rest of genes have identical structure and partial correlations. For scenario (ii), all genes are connected with the same structures under the two conditions but with varying degrees of partial correlations. For scenario (iii), we set both the precision matrix structures and weights differently. Each of the above scenarios is simulated for 100 times.

To mimic the real scRNA-seq data, we fit the ZIP model to the real single-cell dataset (Hovestadt et al., 2019) using GAMLSS (Rigby and Stasinopoulos, 2005), and the result is shown in Web Figure 2. The estimated parameters from the model are applied to the prespecified parameters used in steps (a)–(d) for the simulation study—we simulate $\theta_i$ from Gamma(1.5, 0.1) and $p_i$ from Beta(3, 1). Next, we benchmark both the data processing methods and the tuning parameter selection criteria. Benchmark metrics employed in this simulation study include: Pearson correlation and sum of squared errors (SSEs) between estimated and the true precision matrices. In addition, area under the receiver operating characteristic (AUROC) curve and area under the precision recall (AUPRC) curve are calculated and compared. The data processing methods include the following: (i) NoDropout—there is no dropout, and we consider this as the best performance that can be achieved; (ii) Hybrid—read counts are continuized and imputed by the Bayesian ZIP model and by integrating the iterative McImpute procedure; (iii) Bayesian—read counts are continuized and imputed by the Bayesian ZIP model only; (iv) McImpute—impute the dropouts by McImpute (Mongia et al., 2019); and (v) Observed—the observed raw counts without imputation. (vi) In addition, to demonstrate the advantages of JGNsc, we separately construct a GRN for each condition using GENIE3 (Huynh-Thu et al., 2010), one of the top-performing GRN construction approaches for scRNA-seq data under one condition shown by existing benchmark studies (Blencowe et al., 2019; Pratapa et al., 2020).

For joint graphical models, tuning parameter selection criteria include AIC, BIC, EBIC, and StARS. The results show that the impact of $\lambda_1$, which controls the sparsity of the network, is larger than the impact of $\lambda_2$, and that the performance of JGNsc is stable within a reasonable range of $\lambda_2$ values. Web Figure 3 shows that AIC outperforms the other three criteria for JGNsc_Hybrid; we therefore use AIC in the real data analysis. For No Dropout (Web Figure 4), JGNsc_Bayesian (Web Figure 5), McImpute (Web Figure 6), and the Observed (Web Figure 7) scenarios, AIC also outperforms the other three criteria. Overall, our simulation study shows that the proposed Bayesian continuization and imputation procedure is powerful for
FIGURE 2 Benchmark of data processing methods with sample size varying. GENIE3 is applied to each subgroup of data separately. As the sample size increases, the overall performance is improved for each method. The JGNsc_Hybrid method outperforms other benchmark methods in terms of all four metrics for the joint network partial correlation. The performance of JGNsc_Bayesian (without iteration) becomes better and excel McImpute when the sample size is large (500). This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

improving the downstream gene network construction. Simulation results on scenarios (i) (Figure 2 and Web Figure 8), (ii) (Web Figure 9), and (iii) (Web Figure 10) all suggest that joint modeling of GRNs using JGNsc achieves better performance than separate modeling of GRNs using GENIE3. Given that JGNsc is expected to improve estimation accuracy by exploiting shared information between related conditions, in Web Figure 11, we further summarize the ROC curves of JGNsc and GENIE3 in detecting shared edges by the two conditions. For all the three simulated scenarios, JGNsc is more powerful than GENIE3, highlighting its utility on joint network construction.

When analyzing big datasets, JGNsc can be performed in two separate steps. First, we select a larger gene set to perform the hybrid imputation step; then for the pathways of interest, we parallelly select tuning parameters for the JGL model for pathways of interest. The computational cost of JGNsc JGL model part is $O(G^2)$.

3.2 | Analysis of scRNA-seq data of medulloblastoma and glioblastoma

Medulloblastoma (MB) is the most common pediatric brain tumor, for which oncogenic drivers, including specific transcription factors (TFs), have been well defined. Oncoproteins essential in MB tumorigenesis include MYC, a key oncogenic TF for Group 3 MBs (Northcott et al., 2011), and OTX2, a TF functionally interacting with MYC in Group 3 MBs and playing essential oncogenic roles in Group 4 MBs (Lu et al., 2017; Boulay et al., 2017).

As key oncoproteins in a variety of human cancers, MYC and OTX2 are believed to play essential roles in controlling the transcription of a large number genes and regulating multiple cellular processes in cancer cells, including, most prominently, protein biogenesis and cell metabolism (Dang, 2012; Yang et al., 2021). We postulate that JGNsc can help delineate the roles of MYC and OTX2 in regulating
FIGURE 3  JGNsc networks for Medulloblastoma data. In a network, each node is a gene and each edge shows the partial correlation between a pair of genes. If the partial correlation is zero, there is no connection between this pair of genes. Further, if two genes are connected by a red line, then their partial correlation is positive; otherwise, if they are connected by a green line, their partial correlation is negative. (A) Network visualization of JGNsc results for MYC-related genes from purine enzymes metabolism pathway. Connections between non-MYC related genes are not shown. (B) Network visualization of JGNsc results for OTX2-related genes from purine enzymes metabolism pathway. Connections between non-OTX2-related genes are not shown. (C) Gene set enrichment analysis (GSEA) for metabolism enzymes genes connected to MYC or OTX2. The composition of genes refers to the corresponding fraction of metabolic enzyme genes mapped to each of the listed pathways of interest. The three sets include one set of 651 metabolic enzyme genes, as well as two sets of metabolic enzyme genes found by JGNsc that are connected to MYC or OTX2 in either Group 3 or the intermediate group. The three sets are columns labeled as Reference, Group 3, and Intermediate, respectively. As an example, 7% of the genes from Group 3 belong to the Anaplerotic Reactions of the TCA Cycle pathway. In contrast, only 2% of the Reference set are mapped to that pathway. GSEA is performed with Fisher's exact test and p-values less than 0.05 are shown. The complete GSEA results are available in Web Figure 13. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.
gene transcription as well as their functional interaction in different subtypes of MB cells. Utilizing the MB scRNA-seq dataset by Hovestadt et al. (2019), we selected samples from a subset of 17 individuals that were grouped into two subtypes of medulloblastoma (Group 3 and Group 4) based on their molecular profiles (Hovestadt et al., 2019), as well as another subset of cells that falls between these two subtypes (Intermediate group) (Web Figure 12).

We continue to impute the dropout events in the scRNA-seq data using a selected set of ~6K genes (genes of interest and genes with nonzero counts in at least 300 cells in each group), and use the enzyme-related genes from the mammalian metabolic enzyme database (Corcoran et al., 2017) for the joint network inference. AIC is used to select the tuning parameter values for the JGNsc model. We visualize the MYC-connected and OTX2-connected genes under different conditions in Figure 3. For MYC-connected genes (Figure 3A), the network for group 3 is denser compared to the intermediate group, whereas no connection is detected for MYC in the group 4 samples, in agreement with the prominent roles of MYC in Group 3 MBs (Northcott et al., 2011; Hovestadt et al., 2019). Interestingly, the parallel analysis on OTX2 leads to the following two observations: (i) OTX2 is connected to metabolic genes in the intermediate group of MB cells but less so in the Group 4 MB cells; and (ii) in Group 3 MBs, where OTX2 is thought to be functionally cooperating with MYC (Boulay et al., 2017; Bunt et al., 2011), its connection to metabolic genes was distinct from MYCs, highlighting the unique link between MYC and metabolism in Group 3 MB cells (Figure 3B). Further gene set enrichment analysis (GSEA) based on the previously defined metabolic pathways (Corcoran et al., 2017) reveals that overlapping yet distinct metabolic pathways were enriched for the two TFs (Figure 3C and Web Figure 13). Collectively, these results from the JGNsc analysis suggest that the role of MYC in regulating the expression of metabolic genes is MB-subtype-dependent, and that MYC and OTX2 likely play diverse roles in regulating the transcription of metabolic genes.

Next, we apply the JGNsc framework to a scRNA-seq data from another type of brain tumor, GBM. With a median survival of about 12–15 months, GBM is the most common and lethal brain tumor, due to its intratumoral heterogeneity and treatment resistance. It is believed that these features of GBMs are largely due to the existence of a subset of tumor initiating cells—GBM stem cells (GSCs) (Lathia et al., 2015), and that MYC, a TF in regulating the expression of metabolic genes, is essential for sustaining this population of tumor cells (Wang et al., 2008, 2017).

To test the connection between MYC and the expression of metabolic enzyme-encoding genes in GBM tumor cells exhibiting prevalent heterogeneity, we apply JGNsc to a GBM scRNA-seq dataset from Neftel et al. (2019). In agreement with the heterogeneous nature of GBM tumor cells, GBM scRNA-seq samples are classified into four subgroups with distinct gene expression profiles: neural-progenitor-like (NPC-like), oligodendrocyte-progenitor-like (OPC-like), astrocyte-like (AC-like), and mesenchymal-like (MES-like) tumor cells (Neftel et al., 2019). For simplicity, we only take the malignant single-cell samples from adult GBM patients sequenced by the Smart-seq2 protocol (Picelli et al., 2013), and classify these tumor cells into four subgroups as defined by the original study (Web Figure 14). This results in 1213 OPC-like cells, 1267 NPC-like cells, 1262 AC-like cells, and 637 MES-like cells. Our analysis finds that there are over 20% of cells expressing MYC in each of the subgroups of GBM cells (Web Figure 14B, D) in agreement with the essential roles of MYC in maintaining a subset of GSCs (Wang et al., 2008, 2017). Further joint network analysis on MYC’s connection to metabolism enzyme-encoding genes reveals distinct connection profiles in the four subgroups of GBM cells. MES-like tumor cells demonstrate the most connected gene, followed by AC-like and OPC-like cells, whereas only one gene is found to be connected to MYC in the OPC-like subgroup (Web Figure 15). Subsequent GSEA identifies tumor cell subgroup-dependent metabolic pathway’s connection to MYC: Anaplerotic Reactions of the TCA Cycle, Folic Acid Metabolism are enriched in the MES-like group, Lipid Metabolism is upregulated in the OPC-like group, whereas Pyrimidine Metabolism is enriched in the AC-like group (Web Figure 16). Echoing the findings from the MB cells, these results suggest that MYC’s roles in regulating metabolic processes and the mechanisms underlying MYC-mediated GSC maintenance likely vary in different subgroups of GBM cells.

4 | DISCUSSION

We propose an integrated framework JGNsc to construct joint gene networks using scRNA-seq data under different conditions. The JGNsc framework consists of three major steps: first, continuize and impute scRNA-seq data using an MCMC procedure; second, transform the data into Gaussian form using nonparanormal transformation; and third, jointly construct gene network using JGL. The novelty of our paper lies in the MCMC data continuization step for scRNA-seq data, where iterative matrix completion imputation is implemented. This step helps to correct the posterior distribution of the nondropout events, and, in turn, improve the expression estimation of the dropout events.

In our simulation, we also demonstrate that a proper imputation of the scRNA-seq data would greatly improve
the downstream joint network construction performance. In addition, JGNsc performs consistently better than the other benchmark methods when we vary the sample size and data structure. By transforming the estimated precision matrices into the partial correlation matrices, our method also allows the investigation of the relative magnitude and direction of gene-wise connections.

The following penalty function can be further explored:

\[ P(\{\Omega\}) = \lambda_1 \sum_{k=1}^{K} \sum_{u \neq v} |\omega_{uv}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{u,v} \gamma_{uv}^{k,k'} |\omega_{uv}^{(k)} - \omega_{uv}^{(k')}|, \]

where \( \gamma_{uv}^{k,k'} = 1/|\omega_{uv}^{(k)} - \omega_{uv}^{(k')}| \) is the weight assigned for the pair of \( \omega_{uv}^{(k)} \) and \( \omega_{uv}^{(k')} \) to constrain the level of their penalty, and \( \gamma \) is a positive constant for adjustment of the adaptive weight matrix. The values \( \omega_{uv}^{(k)} \) and \( \omega_{uv}^{(k')} \) are the estimated precision matrices elements from \( K \) naive GGMs. The idea is motivated by the adaptive lasso of Zou (2006), which has the oracle properties and allows consistent variable selection, and the condition-adaptive fused graphical lasso for bulk RNA-seq data of Lyu et al. (2018). By imposing a gene-pair-specific binary weight between conditions, Lyu et al. (2018) only put constraints on edges that are tested to be nondifferentially coexpressed. Such strategies could potentially further improve the network constructions.

JGNsc is, by its default, developed for Smart-Seq2-based scRNA-seq data. For UMI data (e.g., by 10X Genomics), some research (Svensson, 2020; Kim et al., 2020) has suggested that they are likely not zero-inflated. However, the proposed framework can be extended for UMI-based data—instead of directly applying JGNsc, we suggest replacing the ZIP hybrid imputation procedure by JGNsc with a simpler Bayesian Poisson model to continuize the data, similar to the work proposed by Jia et al. (2017) for bulk RNA-seq data. We include in the R package for JGNsc an option for only continuizing UMI data, which bypasses the iterative imputation procedure by setting the parameters dropThreshold = 0, and AvoidIterImp = T in the RunJGNsc() function.

Computationally, JGNsc involves two separate steps. First, the hybrid imputation step, which is \( O(n^2) \), is performed on the entire genes. Second, the JGL model, which is \( O(G^2) \), is performed on individual pathways. Here \( n \) and \( G \) are the number of samples and the number of genes in the pathway of interest, respectively. The second JGL step can be parallelly run on multiple pathways. The two real data analyses were performed on a computing cluster, where the general computing nodes have 24 physical cores, 2.50 GHz Intel processors, 30M cache (Model E5-2680 v3), 256-GB RAM, and 2x10Gbps NIC with R version 4.0.4 (64 bit). The hybrid imputation step took around 1.5–4 hours, depending on the sample size of each condition. Given a pair of tuning parameters, the second JGL took only a few minutes for the metabolism enzyme gene set containing 651 genes. As our current focus is on pathway/gene set-specific networks, we further summarize the number of genes associated with the pathways in the Molecular Signatures Database (MSigDB, provided in R package msigdb). MSigDB is one popular database that provides up-to-date pathway information for Homo sapiens and Mus musculus. For both species, 96% and 99% of the pathways in MSigDB contain less than 500 and 1000 genes, respectively, which can be efficiently handled by JGNsc. To model a large number of genes, such as over 10k in Shang et al. (2020), we suggest adapting, for example, the \( \psi \)-learning algorithm (Liang et al., 2015; Jia et al., 2017).

**ACKNOWLEDGMENTS**

The authors thank the editor and reviewers for their careful read and helpful suggestions.

**DATA AVAILABILITY STATEMENT**

The raw data analyzed in Section 3.2 come from Hovestadt et al. (2019) and Neftel et al. (2019), and are available at Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE119926 and GSE131928. The R code and the processed real datasets are available with this paper at the Biometrics website on Wiley Online Library.

**OPEN RESEARCH BADGES**

This article has earned Open Data and Open Materials badges. Data and code are available at https://doi.org/10.7910/DVN/DHLRSI.

**ORCID**

Meichen Dong https://orcid.org/0000-0001-9437-9290

**REFERENCES**

Akaike, H. (1998) Information theory and an extension of the maximum likelihood principle. In: Kitagawa, G.K., Tanabe, K. & Parzen, E. (Eds.) Selected papers of Hirotugu Akaike. Cham: Springer, pp. 199–213.

Blencowe, M., Arneson, D., Ding, J., Chen, Y.-W., Saleem, Z. & Yang, X. (2019) Network modeling of single-cell omics data: challenges, opportunities, and progresses. Emerging Topics in Life Sciences, 3, 379–398.

Boulay, G., Awad, M.E., Riggi, N., Archer, T.C., Iyer, S., Boonseng, W.E. et al. (2017) OTX2 activity at distal regulatory elements shapes the chromatin landscape of group 3 medulloblastoma. Cancer Discovery, 7, 288–301.

Buil, A., Brown, A.A., Lappendalainen, T., Viñuela, A., Davies, M.N., Zheng, H.-F. et al. (2015) Gene-gene and gene-environment
interactions detected by transcriptome sequence analysis in twins. Nature Genetics, 47, 88–91.

Bunt, J., Hasselt, N.E., Zwijnenburg, D.A., Koster, J., Versteeg, R. & Kool, M. (2011) Joint binding of OTX2 and MYC in promoter regions is associated with high gene expression in medulloblastoma. PLoS One, 6, e26058.

Chen, J. & Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95, 759–771.

Chen, S. & Mar, J.C. (2018) Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics, 19, 1–21.

Corcoran, C.C., Grady, C.R., Pisitkun, T., Parulekar, J. & Knepper, M.A. (2017) From 20th century metabolic wall charts to 21st century systems biology: database of mammalian metabolic enzymes. American Journal of Physiology-Renal Physiology, 312, F533–F542.

Danaher, P., Wang, P. & Witten, D.M. (2014) The joint graphical lasso for inverse covariance estimation across multiple classes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 373–397.

Dang, C. V. (2012) MYC on the path to cancer. Cell, 149, 22–35.

Dong, M. & Jiang, Y. (2019) Single-cell allele-specific gene expression analysis. In: Yuan, G.-C. (Ed.) Computational methods for single-cell Data analysis. New York, NY: Springer, pp. 155–174.

Hicks, S.C., Townes, F.W., Teng, M. & Irizarry, R.A. (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics, 19, 562–578.

Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.L., Baumgartner, A. et al. (2019) Resolving medulloblastoma cellular architecture by single-cell genomics. Nature, 572, 74–79.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. & Geurts, P. (2010) Inferring regulatory networks from expression data using tree-based methods. PLoS One, 5, e12776.

Jia, B., Xu, S., Xiao, G., Lamba, V. & Liang, F. (2017) Learning gene regulatory networks from next generation sequencing data. Bioinformatics, 73, 1221–1230.

Jiang, Y., Zhang, N.R. & Li, M. (2017) SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biology, 18, 74.

Karlebach, G. & Shamir, R. (2008) Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology, 9, 770–780.

Kim, T.H., Zhou, X. & Chen, M. (2020) Demystifying “drop-outs” in single-cell UMI data. Genome Biology, 21, 1–19.

Langfelder, P. & Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559.

Lathia, J.D., Mack, S.C., Mulkearns-Hubert, E.E., Valentim, C.L. & Rich, J.N. (2015) Cancer stem cells in glioblastoma. Genes & Development, 29, 1203–1217.

Lee, S., Liang, F., Cai, L. & Xiao, G. (2018) A two-stage approach of gene network analysis for high-dimensional heterogeneous data. Biostatistics, 19, 216–232.

Liang, F., Song, Q. & Qiu, P. (2015) An equivalent measure of partial correlation coefficients for high-dimensional Gaussian graphical models. Journal of the American Statistical Association, 110, 1248–1265.

Liu, H., Lafferty, J. & Wasserman, L. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research, 10, 2295–2328.

Liu, H., Roeder, K. & Wasserman, L. (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. In: Advances in Neural Information Processing Systems, pp. 1432–1440.

Lu, Y., Labak, C.M., Jain, N., Purvis, I.J., Guda, M.R., Bach, S.E. et al. (2017) OTX2 expression contributes to proliferation and progression in MYC-amplified medulloblastoma. American Journal of Cancer Research, 7, 647–656.

Lyu, Y., Xue, L., Zhang, F., Koch, H., Saba, L., Kechriss, K. & Li, Q. (2018) Condition-adaptive fused graphical lasso (CFGL): an adaptive procedure for inferring condition-specific gene co-expression network. PLoS Computational Biology, 14, e1006436.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M. et al. (2012) Wisdom of crowds for robust gene network inference. Nature Methods, 9, 796–804.

Mongia, A., Sengupta, D. & Majumdar, A. (2019) MclImpute: Matrix completion based imputation for single cell RNA-seq data. Frontiers in Genetics, 10, 9.

Nefelt, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J. et al. (2019) An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell, 178, 835–849.

Northcott, P.A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C.G., Mack, S. et al. (2011) Medulloblastoma comprises four distinct molecular variants. Journal of Clinical Oncology, 29, 1408–1414.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G. & Sandberg, R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods, 10, 1096–1098.

Pratapa, A., Jalilah, A.P., Law, J.N., Bharadwaj, A. & Murali, T. (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods, 17, 147–154.

Rigby, R.A. & Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape, (with discussion). Applied Statistics, 54, 507–554.

Schwarz, G. (1978) Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

Shang, L., Smith, J.A. & Zhou, X. (2020) Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS Genetics, 16, e1008734.

Sun, Y., Babu, P. & Palomar, D.P. (2016) Majorization-minimization algorithms in signal processing, communications, and machine learning. IEEE Transactions on Signal Processing, 65, 794–816.

Svensson, V. (2020) Droplet scRNA-seq is not zero-inflated. Nature Biotechnology, 38, 147–150.

Townes, F.W., Hicks, S.C., Aryee, M.I. & Irizarry, R.A. (2019) Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biology, 20, 1–16.

Trappnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M. et al. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32, 381–386.

Van de Sande, B., Flerin, C., Davie, K., De Waegeneer, M., Hulselmans, G., Albar, S. et al. (2020) A scalable scenic workflow for single-cell gene regulatory network analysis. Nature Protocols, 15, 2247–2276.
Wang, J., Wang, H., Li, Z., Wu, Q., Lathia, J.D., McLendon, R.E. et al. (2008) c-MYC is required for maintenance of glioma cancer stem cells. *PLoS One*, 3, e3769.

Wang, X., Yang, K., Xie, Q., Wu, Q., Mack, S.C., Shi, Y. et al. (2017) Purine synthesis promotes maintenance of brain tumor initiating cells in glioma. *Nature Neuroscience*, 20, 661–673.

Wu, N., Yin, F., Ou-Yang, L., Zhu, Z. & Xie, W. (2020) Joint learning of multiple gene networks from single-cell gene expression data. *Computational and Structural Biotechnology Journal*, 18, 2583–2595.

Wysocki, A.C. & Rhemtulla, M. (2019) On penalty parameter selection for estimating network models. *Multivariate Behavioral Research*, 56, 288–302.

Yahav, I. & Shmueli, G. (2012) On generating multivariate poisson data in management science applications. *Applied Stochastic Models in Business and Industry*, 28, 91–102.

Yang, R., Wang, W., Dong, M., Roso, K., Greer, P., Bao, X. et al. (2021) Distribution and vulnerability of transcriptional outputs across the genome in MYC-amplified medulloblastoma cells. *bioRxiv*.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

**SUPPORTING INFORMATION**
Web Appendices and Figures referenced in Sections 2 and Section 3, R code, and processed real datasets are available with this paper at the Biometrics website on Wiley Online Library. R package for JGNsc is also available online at github: https://github.com/meichendong/JGNsc.

**How to cite this article:** Dong, M., He, Y., Jiang, Y., Zou, F. (2023) Joint gene network construction by single-cell RNA sequencing data. *Biometrics*, 79:915–925. https://doi.org/10.1111/biom.13645