Single-cell RNA-seq Imputation using Generative Adversarial Networks

Yungang Xu[1,†], Zhigang Zhang[2,6,†], Lei You[1], Jiajia Liu[1,4], Zhiwei Fan[1,5], Xiaobo Zhou[1,3,*]

[1] Center for Computational Systems Medicine, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX 77030, USA

[2] School of Information Management and Statistics, Hubei University of Economics, Wuhan, Hubei 430205, China

[3] Department of Pediatric Surgery, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[4] School of Electronics and Information, Tongji University, Shanghai, Shanghai 201804, China

[5] West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, Chengdu 610040, China

[6] Hubei Center for Data and Analysis, Hubei University of Economics, Wuhan, Hubei, 430205, China

[†]These authors contributed equally to this work.

[*]Correspondence: xiaobo.zhou@uth.tmc.edu

Email addresses:

YX: yungangx.xu@uth.tmc.edu

ZZ: zzg@hbue.edu.cn

LY: lei.you@uth.tmc.edu

JL: jiajia.liu@uth.tmc.edu

ZF: zhiwei.fan@uth.tmc.edu

XZ: xiaobo.zhou@uth.tmc.edu

Abstract

Single-cell RNA-seq (scRNA-seq) enables the characterization of transcriptomic profiles at the single-cell resolution with increasingly high throughput. However, it suffers from many sources of technical noises, including insufficient mRNA molecules that lead to excess false zero values, often termed dropouts. Computational approaches have been proposed to recover the biologically meaningful expression by borrowing information from similar cells in the observed dataset. However, these methods suffer oversmoothing and removal of natural cell-to-cell stochasticity in gene expression. Here, we propose the generative adversarial networks for scRNA-seq imputation (scIGANs), which uses generated realistic rather than observed cells to avoid these limitations and the powerless for rare cells. Evaluations based on a variety of simulated and real scRNA-seq datasets demonstrate that scIGANs is effective for dropout imputation and enhancing various downstream analysis. ScIGANs is also scalable and robust to small datasets that have few genes with low expression and/or cell-to-cell variance.

Introduction

Single-cell RNA-seq (scRNA-seq) revolutionizes the traditional profiling of gene expression, making it able to fully characterize the transcriptomes of individual cells at the unprecedented throughput. A major problem for scRNA-seq is the sparsity of the expression matrix with a tremendous number of zero values. Most of these zero or near-zero values are artificially caused by technical defects including but not limited to low capture rate, insufficient sequencing depth, or other technological factors such that the observed zero does not reflect the underlying true expression level, which is called dropout [1]. A pressing need in scRNA-seq data analysis remains identifying and handling the dropout events that, otherwise, will severely hinder downstream analysis and attenuate the power of scRNA-seq on a wide range of biological and biomedical applications. Therefore, applying computational approaches to address problems of

missingness and noises is very important and timely, particularly considering the increasingly popular and large amount of scRNA-seq data.

Several methods have been recently proposed to address the challenges resulted from excess zero values in scRNA-seq. MAGIC [2] imputes missing expression values by sharing information across similar cells, based on the idea of heat diffusion. ScImpute [3] learns each gene's dropout probability in each cell and then imputes the dropout values borrowing information from other similar cells selected based on the genes unlikely affected by dropout events. SAVER [4] borrows information across genes using a Bayesian approach to estimate unobserved true expression levels of genes. DrImpute [5] impute dropouts by simply averaging the expression values of similar cells defined by clustering. VIPER [6] borrows information from a sparse set of local neighborhood cells of similar expression patterns to impute the expression measurements in the cells of interest based on nonnegative sparse regression models. Meanwhile, some other methods aim at the same goal by denoising the scRNA-seq data. DCA [7] uses a deep count autoencoder network to denoise scRNA-seq datasets by learning the count distribution, overdispersion, and sparsity of the data. ENHANCE [8] recovers denoised expression values based on principal component analysis on raw scRNA-seq data. During the preparation of this manuscript, we also noticed another imputation method DeepImpute [9], which uses a deep neural network with dropout layers and loss functions to learn patterns in the data, allowing for scRNA-seq imputation.

While existing studies have adopted varying approaches for dropout imputation and yielded promising results, they either borrow information from similar cells or aggregate (co-expressed or similar) genes of the observed data, which will lead to oversmoothing (e.g. MAGIC) and remove natural cell-to-cell stochasticity in gene expression (e.g. scImpute). Moreover, the imputation performance will be significantly reduced for rare cells, which have limited information and are common for many scRNA-seq studies. Alternatively, SCRABBLE [10]

attempt to leverage bulk data as a constraint on matrix regularization to impute dropout events. However, most scRNA-seq studies often lack matched bulk RNA-seq data and thus limit its practicality. Additionally, due to the non-trivial distinction between true and false zero counts, imputation and denoising need account for both the intra-cell-type dependence and inter-cell-type specificity. In view of the above concerns, a deep generative model would be a better choice to learn the true data distribution and then generate new data points with some variations, which are then independently used to impute the missing values and avoid overfitting.

Deep generative models have been widely used for missing value imputation in fields [11-13], however, other than scRNA-seq. Although a deep generative model was used for scRNA-seq analysis [14], it's not explicitly designed for dropout imputation. Among deep generative models, generative adversarial networks (GANs) have evoked increasing interest in the computer vision community since its first introduction in 2014 [15]. GANs has become an active area of research with multiple variants developed [16-20] and holds promising in data imputation [21] because of its capability of learning and mimicking any distribution of data. Given the great success of GANs in inpainting, we hypothesize that similar deep neural net architectures could be used to impute dropouts in scRNA-seq data.

In this study, we propose a GANs framework for scRNA-seq imputation (scIGANs). Inspired by its established applications in inpainting, we convert the expression profile of each individual cell to an image, wherein the pixels are represented by the normalized gene expression. And then dropout imputation becomes the process of inpainting an image by recovering the missing pieces that represent the dropout events. Because of the inherent advantages of GANs, scIGANs does not impose an assumption of specific statistical distributions for gene expression levels and dropout probabilities. It also does not force the imputation of genes that are not affected by dropout events. Moreover, scIGANs generates a set of realistic single cells instead of directly borrowing information from observed cells to impute the dropout events, which can

avoid overfitting for the cell type of big population and meanwhile promise enough imputation power for rare cells. Using a variety of simulated and real datasets, we extensively evaluate scIGANs with nine other state-of-the-art, representative methods and demonstrate its superior performance in recovering the biologically meaningful expression, identifying subcellular states of the same cell types, improving differential expression and temporal dynamics analysis. ScIGANs is also robust and scalable to datasets that have a small number of genes with low expression and cell-to-cell variance.

Results

1. The scIGANs approach

Generative adversarial networks (GANs), first introduced in 2014 [15], evoked much interest in the computer vision community and has become an active area of research with multiple variants developed [16-20]. Inspired by its excellent performance in generating realistic images [22-26] and recent application to generating realistic scRNA-seq data [27, 28], we propose scIGANs, the generative adversarial networks for scRNA-seq imputation (Figure 1, Methods). The basic idea is that scIGANs can learn the non-linear gene-gene dependencies from complex, multi-cell type samples and train a generative model to generate realistic expression profiles of defined cell types [27, 28]. To train scIGANs, the real single-cell expression profiles are first reshaped to images and fed to GANs, wherein each cell corresponds to an image with the normalized gene expression representing the pixel (Figures 1 and S1A, Methods). The generator generates fake images by transforming a 100-dimensional latent variable into single-cell gene expression profiles (Figure S1A). The discriminator evaluates whether the images are authentic or generated. These two networks are trained concurrently whilst competing against one another to improve the performance of both (Figure 1).

Once trained, the generative model is used to generate scRNA-seq data of defined cell types. And then we propose to infer the true expression of dropouts from the generated realistic cells.

The most important benefit of using generated cells instead of the real cells for scRNA-seq imputation is to avoid overfitting for the cell type of big population but insufficient power for rare cells. The generator can produce a set of cells of any number with the expression profiles faithfully characterizing the demand cell type; and then the k-nearest neighbors (KNN) approach is used to impute the dropouts of the same cell type in the real scRNA-seq data (Figure S1B, Methods). The scIGANs is implemented in python and R, and compiled as a command-line tool compatible with both CPU and GPU platform. The core model is built on the PyTorch framework and adopted to accommodate scRNA-seq data as input. It's publicly available at https://github.com/xuyungang/scIGANs.

2. ScIGANs recovers single-cell gene expression from dropouts without inflicting extra noise

Recovery of the biologically meaningful expression from dropout events is the major goal of scRNA-seq imputation to benefit the downstream analyses and biological discoveries. We use both simulated and real scRNA-seq datasets to illustrate the performance and robustness of scIGANs in rescuing dropouts and avoiding additional noise from imputation.

First, simulated datasets are used to evaluate the imputation performance since they have known 'truth' and can thus benchmark different methods. In a single dataset with a 52.8% zero rate that was simulated according to an independent single-cell clustering method CIDR [29] (Methods), scIGANs performed superiorly over all other nine methods in recovering the gene expression and cell population clusters (Figures 2A and S2A; Tables S1). Although GANs is a supervised model that requires pre-defined cell labels, we implemented scIGANs to accommodate scRNA-seq data without prior labels, instead to learn the labels by applying spectral clustering [30] on input data. On the same simulated data, scIGANs trained without labels (scIGANs w/o) reduced the performance slightly and remained the superiority over the other eight compared methods, except for scImpute [3] (Figures 2A and S2A; Tables S1).

Second, we test the performance of scIGANs and other peer methods on datasets with different dropout rates simulated by Splatter [31] (Methods). scIGANs ranks in the top in rescuing the population clusters (Figures S2B-D) and has the highest resistance to dropout rate increase (Figure S2E; Table S2). Moreover, to evaluate the robustness of imputation methods, we used the same simulation strategy described by SCRABBLE [32] to repeat the above Splatter simulation 100 times for each dropout rate. We evaluated the performance by multiple quantitative clustering metrics (Table S3). The second-ranked SCRABBLE performed superiorly over all other peer methods, however, it has worse concordance among simulated replicates with a higher dropout rate (Figure 2B). In contrast, scIGANs ranks in top among all methods and has the most robust performance among the replicates across increasing dropout rates (Figures 2B and S3A-F; Tables S3).

Third, we evaluate the imputation methods using real scRNA-seq data from the Human brain, which contains 420 cells in eight well-defined cell types after we excluded uncertain hybrid cells [33] (Methods). However, the raw data doesn't show clear clustering of all cell types because of the dropouts and technical noise. After imputation, scIGANs enhanced the cell type clusters to the maximum extent so that all 8 cell types could be separated and identified (Figure 2C). Quantitative evaluations of the clustering following different imputation methods highlighted the superiority of scIGANs over the others, even trained without the prior cell labels (Figures 2D and S3G; Table S4).

Last, we test another important yet difficult to quantify robustness, i.e. to what extent the imputation method will not introduce extra noise by, for example, mistakenly imputing biological "zeros" or over-imputation. None of the existing imputation methods evaluated their robustness in avoiding extra noise using real scRNA-seq data. Spike-in RNA (e.g. ERCC spike-in developed by the External RNA Controls Consortium) is a common set of external RNA controls to be equally added to an RNA analysis experiment after sample isolation. It is widely used in scRNA-seq experiments to remove the confounding noises from biological variance. Because

the spike-in RNAs are added to samples with the identical amount to capture the technical noise, the readout for the spike-in RNAs should be free of cell-to-cell variability and the detected variances of expression, if exists, should only come from technical confounders other than biological contexts (e.g. cell types). Therefore, the expression of spike-in RNAs that were added to individual cells should not be able to cluster these cells into different subgroups regarding cell types or other biological states. We here use the ERCC spike-in read counts from a real scRNA-seq study [34] to evaluate the imputation methods on denoising the technical variance without introducing extra noise (Methods). These 92 ERCC RNAs were added to 288 single-cell libraries of three sets of 96 cells with different cell-cycle states. However, the raw counts failed to cluster these cells into one cluster due to the dropouts of spike-in RNAs (Figure 2E). We expected that the imputation could help impute the artificial zeros without exposing the cell states to spike-in profiles and thus all cells should have the same spike-in profiles and will be clustered into a single group. ScIGANs successfully recovers the spike-in profiles with minimum cell-to-cell variability and clustered all cells closely into one group, even though it was trained with supervisory cell labels (Figures 2E and S3H-I). However, other imputation methods suffer from introducing extra noises and thus made clustering even worse (Figures S3H-I; Table S5). Altogether, scIGANs performs superiorly on imputing the dropouts and avoiding extra noise.

3. ScIGANs enables the identification of cellular states of the same cell type

Single-cell RNA-seq is typically used to identify different cell types from heterogeneous tissues or cell populations. However, cell populations that seem homogeneous, in terms of expression of cell surface markers, comprise many different cellular states and hide cell-to-cell variability that can have significant effects on cell function [35, 36], such as cellular functions, developmental stages, cell cycle phase, and adjacent microenvironments. Therefore, many biological questions require deeper investigation beyond the cell types towards implied cellular states, such as cell-cycle phases of the same cell type. It was reported that cell cycles contribute to phenotypic and functional cell heterogeneity even in monoclonal cell lines [37-39].

However, identifying the different cell-cycle phases of the same cell type from scRNA-seq data is more challenging due to the prevalence of dropout and high technical variance, which was recently reported more attributable than cell cycle to the single-cell transcriptomic variability [38]. We thereby test how imputation could benefit the identification of cell cycle variability from scRNA-seq studies.

First, we reanalyze scRNA-seq data from mouse embryonic stem cells (mESC) that were sorted for G1, S and G2M phases of the cell cycle (Methods) [34]. Due to the dropout and other technical noise, the raw data does not show cluster structures regarding the three different cell-cycle phases (Figure 3A) and has the poorest clustering measurements (Figure S4A). All other imputation methods fail to recover the cluster structure regarding the cell-cycle states (Figures 3A and S4A). Only scIGANs shows significant improvement in detecting cell-cycle states with the best performance (Figures 3A and S4A). Using a collection of independently predefined cell-cycle marker genes (Methods), scIGANs significantly improves the identification of the cell cycle states superior over all other methods, shown as the most of sorted cells are correctly assigned in the cell-cycle phase spaces (Figures 3B and S4B).

Second, we assess the performance of different imputation methods on pinpointing the cell-cycle dynamics using a large scRNA-seq data of about 6.8k mouse ESCs (Methods) [40]. The previous work confirmed that ES cells lack strong cell-cycle oscillations in mRNA abundance, but they do show evidence of limited G2/M phase-specific transcription [40]. Imputation by scIGANs significantly improved the cell-cycle oscillations with especially a more obvious G2/M phase-specific transcription (Figures 3C and S4C-L). All the above demonstrate that scIGANs performs better than all other methods on recovering and capturing the cellular states and very subtle cell-cycle oscillations among single cells.

4. ScIGANs improves the differential expression analysis

Differential expression analysis refers broadly to the task of identifying those genes with expression levels that depend on some variables, like cell type or state. Ultimately, most single-cell studies start with identifying cell populations and characterizing genes that determine cell types and drive them different from one to another. Using the scRNA-seq data [41] that have matched bulk RNA-seq data, we compare the performances of different imputation methods on improving identification of differentially expressed genes (DEGs). This dataset has six samples of bulk RNA-seq (four for H1 ESC and two for definitive endoderm cells, DEC) and 350 samples of scRNA-seq (212 for H1 ESC and 138 for DEC) (Methods). DESeq2 [42] is used to identify DEGs for both bulk and single-cell RNA-seq data between the H1 and DEC cells (Methods). The raw scRNA-seq has a much higher zero expression rate than bulk RNA-seq (49.1% vs 14.8%) and shares fewest DEGs with bulk samples (Figure 4A). After imputation, the number of DEGs is increased toward the DEGs numbers of bulk samples (except the two other neural network-based methods, DCA [7] and DeepImpute [9], which give fewer DEGs than raw data). scIGANs imputation identifies the highest number of dataset-specific DEGs and shares a significant number of DEGs with bulk RNA-seq (Figure 4A). Using a set of top 1000 DEGs from bulk samples (500 up-regulated and 500 down-regulated genes) as a benchmark, scIGANs-imputed scRNA-seq data show the highest correspondence with bulk RNA-seq (Figures 4B and S5).

Moreover, the expressions of five marker genes for H1 and DEC, respectively, were investigated to compare the extent to which the imputation could recover the expression patterns of signature genes. Results show that scIGANs best reflect the expression signatures of both H1 and DEC cells by removing undesirable variation resulted from dropouts (Figures 3C and S6). Projection of cells to the UMAP space overlaid by the expression of signature genes furtherly highlights the performance of scIGANs on recovering the expression patterns of

signature genes (Figures 3D-E and S7). In summary, scIGANs improves the identification of DEGs from scRNA-seq data with better performance.

5. ScIGANs enhances the inference of cellular trajectory

Beyond characterizing cells by types, scRNA-seq also largely benefits organizing cells by temporal or developmental stages, i.e. cellular trajectory. In general, trajectory analysis starts with reducing the dimensionality of the expression data, then reconstructs a trajectory along which the cells are presumed to travel, and finally projects each cell onto this trajectory at the proper position. Although single-cell experiments can illuminate trajectories in a wide variety of biological settings [43-46], none of the single-cell trajectory inference methods account for dropout events. We hypothesized that inferring the cellular trajectory on scRNA-seq data after imputation could improve the accuracy of pseudotime ordering. We utilize a time-course scRNA-seq data derived from the differentiation from H1 ESC to definitive endoderm cells (DEC) [41]. A total of 158 cells were profiled at 0, 12, 24, 36, 72, and 96 hours after inducing the differentiation from H1 ESCs (Figure 5A). We apply scIGANs and all other nine imputation methods to the raw scRNA-seq data with known time points and then reconstruct the trajectories. Imputation by scIGANs produces the highest correspondence between the inferred pseudotime and real-time course (Figures 5B-C and S8), suggesting that scIGANs recovers more accurate transcriptome dynamics along the time course. We also study the signature genes of pluripotency (e.g. NANOG and POU5F1) and DECs (e.g. CER1 and HNF1B) and find that scIGANs improves the gene expression temporal dynamics after imputation (Figures 5D-E) and has better performance than other imputation methods (Figure S8). These results demonstrate that scIGANs can help to improve the single-cell trajectory analysis and recover the temporal dynamics of gene expression.

6. scIGANs is robust to the small dataset of few genes with low expression or cell-to-cell variance

In general, other imputation methods (e.g. SAVER [4] and scImpute [3]) heavily rely on a set of pre-selected informative genes that are highly expressed and unlikely to suffer from the dropout. Imputation is then performed from the most similar cells defined by these informative genes. In contrast, scIGANs automatically learns the gene-gene and cell-cell dependencies from the whole dataset. More important, scIGANs converts each single-cell expression profile to an image so that a 1-dimension "feature" vector is reshaped to a 2-dimension matrix with each element representing the expression of a single gene (Figure S1A). Like image processing, scIGANs is then trained by convolution on the matrix so that the 2-dimension gene-gene relations within each individual cell are captured. Therefore, we hypothesize that scIGANs is more robust to genes of low expression or with less cell-to-cell variance.

From the aforementioned scRNA-seq data with 350 cells (212 H1 ESC and 138 DEC) [41], we randomly sample small sets of genes (n=1024 for each) from the 5000-gene sets with, respectively, top/lower means or variances, as well as a set of 1024 genes randomly from all expressed genes (refer to Methods for details). When visualized only on the 1024 genes with very low expression or variance, the two types of cells are almost mixed up without any cluster characterization for the raw expression profiles (Figures 6A and 4D). Imputation by scIGANs successfully recovered the two cell clusters for both datasets with only 1024 genes of low expression and variance, respectively (Figure 6B). However, all other methods failed in identifying the two cell types from these datasets (Figure S9). Moreover, scIGANs significantly changes the mean and variance of expression after imputation, while it's not always the same cases for other methods (Figures 6C-D and S9). All these results show that scIGANs is robust to a small dataset of genes with very low expression or cell-to-cell variance, which are less informative for other imputation methods. It's strong support to the expectation that scIGANs

can learn very limited gene-gene and cell-cell dependencies from a small set of lowly or close-to-uniform expressed genes.

Discussion

Here we propose the generative adversarial networks for scRNA-seq imputation (scIGANs). ScIGANs converts the expression profiles of individual cells to images and feeds them to generative adversarial networks. The trained generative network produces expression profiles representing the realistic cells of defined types. The generated cells, rather than the observed cells, are then used to impute the dropouts of the real cells. We assess scIGANs regarding its performances on the recovery of gene expression and various downstream applications using simulated and real scRNA-seq datasets. We provide compelling evidence that scIGANs performs superior over the other nine peer imputation methods. Most importantly, using generated rather than observed cells, scIGANs avoids overfitting for the cell type of big population and meanwhile promise enough imputation power for rare cells.

While there are many methods for scRNA-seq imputation, we specifically show how the GANs can improve the imputation and downstream applications, representing one of three pioneering applications of GANs to genomic data. Two other recent manuscripts used GANs to simulate (generate) realistic scRNA-seq data with the applications of either integrating multiple scRNA-seq datasets [19] or augmenting the sparse and underrepresented cell populations in scRNA-seq data [27, 28]. We, for the first time, advance the applications of GANs to scRNA-seq for dropout imputation. Inspired by the great success of GANs in inpainting and a highly relevant work that applied GANs for 'realistic' generation of scRNA-seq data [27, 28], we speculate that the generated realistic cells can not only augment the observed dataset but also benefit the dropout imputation since it was proved that the generated data mimics the distribution of the real data in their original space with stable fidelity [27, 28]. Our multiple downstream assessments and applications on simulated and real scRNA-seq datasets demonstrated its

advantage in dropout imputation, superior over other peer methods. Especially for cells coming from very small populations, generated data were proved to faithfully augment the sparse cell populations [27, 28] and thus reduce the sampling bias and improve the imputation power, which, however, are suffered by all other imputation methods. Additionally, GANs is able to learn dependencies between genes beyond pairwise correlations [27, 28], which enables scIGANs more sensitive and robust to small datasets with very low or uniform expressions. We demonstrated these advantages by ERCC spike-in RNAs (Figures 2E and S3H-I) and downsampling real scRNA-seq data (Figures 6 and S9).

The underlying basis of scIGANs is that the real scRNA-seq data is derived from sampling, which doesn't have enough cells to characterize the true expression profiling of each cell type, even for the major type of the cell populations; and the generated realistic cells could augment the observations, especially for sparse and underrepresented cell populations, and thus improve the dropout imputation of scRNA-seq data. There are many benefits of using realistic rather than the observed cells for imputation. First, the generated cells characterize the expression profiles of real cells and faithfully represent the cell heterogeneity. Therefore, the realistic cells are ideal to serve as extra samples and independently impute the observed dropouts to avoid the “circular logic” issue (overfitting) suffered by other methods (e.g. scImpute), which borrows information from the observed data per se. Second, the realistic cells will augment the rare cell-types, and thus overcome potential sampling biases present in downstream analyses. Additionally, benefitting from the power of GANs in adversarially discriminating between real and realistic data, and the augmentation from generated data, scIGANs is more sensitive to subcellular states like the cell-cycle phases investigated in this study. Imputation by scIGANs enables the investigation of scRNA-seq data beyond the identification and characterization of cell types but go deeper into subcellular states and capture cell-to-cell variability of the homogenous cell populations. This is critical for the applications of scRNA-seq to pinpoint the state transitions along the cellular trajectory or identify and remove the subcellular confounding

factors (e.g. cell-cycle phases) [38]. Our evaluations on cell-cycle phase detection and trajectory construction show the superiority of scIGANs over the all other nine tested methods.

In summary, scIGANs is a method that takes advantage of both the gene-to-gene and cell-to-cell relationships to recover the true expression level of each gene in each cell, removing technical variation without compromising biological variabilities across cells. ScIGANs is also compatible with other single-cell analysis methods since it does not change the dimension (i.e., the number of genes and cells) of the input data and it effectively recovers the dropouts without affecting the non-dropout expressions. Additionally, ScIGANs is scalable and robust to small datasets that have few genes with low expression and/or cell-to-cell variance.

Methods

Generative adversarial networks and improved Wasserstein GANs

We here show that the generative adversarial networks (GANs) can be applied to scRNA-seq imputation. The GANs training strategy is to define a game between two competing networks. The generator network maps a source of noise to the input space. The discriminator network receives either a generated sample or a true data sample and must distinguish between the two. The generator is trained to fool the discriminator. Formally, the game between the generator G and discriminator D is the minimax objective $\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}[\log(D(x))] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[\log(1 - D(\tilde{x}))]$; where D is the discriminator that can be any network, $\mathbb{P}_r$ is the real data distribution and $\mathbb{P}_g$ is the model distribution implicitly defined by $\tilde{x} = G(z), z \sim p(z)$; G is the generator which can be any network, z can be sampled from any noise distribution p, such as the uniform distribution or a spherical Gaussian distribution.

It is difficult to train the original GANs model since minimizing the objective function corresponds to minimizing the Jensen-Shannon divergence between $\mathbb{P}_r$ and $\mathbb{P}_g$, which is not continuous with respect to the generator's parameters. Earth-Mover (Wasserstein-1) distance $W(q, p)$ is used to deal with such difficulty [47]. Such a model is called Wasserstein GANs(WGANs) which the

objective function is constructed as $\min_G \max_{D\in\mathcal{D}} \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})]$; where $\mathcal{D}$ is the set of 1-Lipschitz function, the definition of other symbols are the same as the original GANs model. To enforce the Lipschitz constraint on the critic, one can clip the weights of the critic to lie within a compact space $[-c, c]$. The set of functions satisfying this constraint is a subset of the k-Lipschitz functions for some k which depends on c and the critic architecture. Researchers introduced an alternative way to enforce the Lipschitz constraint, usually called improved WGANs(IWGANs), which is widely used in training GANs models [48]. The objective is $\min_G \max_{D\in\mathcal{D}} \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[D(x)] + \lambda\, \mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]$; where $\hat{x}$ is sampled from the straight lines between pairs of points sampled from the real data distribution and the generator distribution. λ is a predefined parameter. BEGAN [49] is an equilibrium enforcing method paired with a loss derived from the Wasserstein [20] distance for training auto-encoder based Generative Adversarial networks. The BEGAN objective is:

$$\begin{cases} L_D = L(x) - k_t L\big(G(z_D)\big) & \text{for } \theta_D \\ L_G = L\big(G(z_D)\big) & \text{for } \theta_G \\ k_{t+1} = k_t + \lambda_k \Big(\gamma L(x) - L\big(G(z_D)\big)\Big) & \text{for each training step } t \end{cases}$$

where

$$L(v) = |v - D(v)|^{\eta} \text{ where } \begin{cases} D: \mathbb{R}^{N_x} \mapsto \mathbb{R}^{N_x} & \text{is the autoencoder function.} \\ \eta \in \{1,2\} & \text{is the target norm.} \\ v \in \mathbb{R}^{N_x} & \text{is a sample of dimension } N_x \end{cases}$$

In this paper, we use this method to train our scIGANs.

The scIGANs

Although, scIGANs is designed scalable to the dataset with any number of cell types and genes, we here taking a dataset with 9 cell types and 32*32=1024 genes as an example to elucidate how it works. The generator network of scIGANs is defined as $G(z, La_z; \theta)$. The inputs of the generator are: $z \sim \mathrm{norm}(0,1)$, and label $La_z \sim U(1,9)$ (Supplementary Figure S1A). Denote θ as

the parameters need to be learned. To be noted that. The generator is defined as following the steps:

1. Do transposed convolution on z by *GConv1_1* and get the *tensor* z_n of dimension (32,32,32).
2. Do transposed convolution on La_z by *GConv1_2* and get the *tensor* La_n of dimension (8,32,32).
3. Concatenate z_n and La_n to get *GConcat1.*
4. Do convolution on *GConcat1* by *GConv2_1* and *GConv2_2* to get the tensor of dimension (1,32,32), which is the output of the Generator.

The discriminator network is defined as $D(x, La_x; w)$. The inputs of discriminator are samples of real data $x \sim \mathbb{P}_r(\text{or } \tilde{x} \sim \mathbb{P}_g)$ representing the expression profile of an individual cell, and label of $x(\text{or } \tilde{x})$ denoted by La_x representing the cell type or subpopulation. Denote w as the parameters need to be learned. The discriminator is defined as following the steps (supplementary Figure S1A):

1. Do convolution on x or $\tilde{x}$ by *DConv1_1* and get the *tensor* of dimension (16,32,32).
2. Do convolution on La_x by *DConv1_2* and get the tensor of dimension (16,32,32).
3. Concatenate results of steps (1) and (2) as *Dconcat1*, which is a tensor of the dimension (32,32,32).
4. Convert the *Dconcat1* to a vector of length 16 using a fully connected network (FCN).
5. Do convolution on the result of step (4) by *GConv2_1* and *GConv2_2* to get the tensor of dimension (1,32,32), which is the output of the Discriminator.

With a well-trained GANs model, for a given cell c_i which belongs to the subpopulation K_{c_i}, we generate a candidate set $A^{K_{c_i}}$ with n_{can} expression profiles. Denote $c'_{i_{knn}}$ as the k nearest

neighbors using Euclidian distance in the set $A^{K_{c_i}}$. We then use the following equation to impute jth gene in the cell c_i (Supplementary Figure S1B): $\hat{c}_{i,j} = \begin{cases} c_{i,j}, & if\ c_{i,j} > 0 \\ c'_{i_{knn},j}, & else \end{cases}$.

Data processing and normalization

The data of a scRNA-seq study are usually organized as a read count matrix with N rows representing genes and M columns representing cells, which is the input of scIGANs. Since scIGANs is trained similarly to the training for image processing, we need to transfer the expression profile of each cell to a grayscale image (Supplementary Figure S1A). To this end, scIGANs firstly normalizes the raw count matrix by the maximum read count of each sample (cell) so that all genes of each sample will have the expression values in a [0,1] range. scIGANs then reshapes the expression profile of each cell to a square image in a column-wise manner, with the normalized gene expression values representing the pixels of the image. The image size will be $n \times n$, where n is the minimum integer so that $n \times n \geq \mathrm{N}$. If the gene number is less than $n \times n$, extra zeroes will be filled. Then, a scRNA-seq matrix with M cells will be represented as M grayscale images and used to train a conditional GANs with the cell labels.

Simulated scRNA-seq data

We first simulated a simple scRNA-seq data with 150 cells and 20180 genes using the default CIDR simulation function scSimulator(N=3, k=50) [29]. Three cell types are generated with 50 cells for each. The dropout data has a dropout rate of 52.8%. Figures 2A, S2A and Table S1 are derived from this data. We then tested the performance of different imputation methods on different dropout rates simulated by Splatter [31]. We took the same simulation strategy with the same parameters as the Splatter simulator used by SCRABBLE [10]. Specifically, three scRNA-seq datasets with three different dropout rates (71%, 83%, and 87%) were simulated; each dataset has 800 genes and 1000 cells grouped into three clusters (cell types). Figures S2B-E and Table S2 were derived from these datasets. To test the robustness of imputation methods,

we repeated 100 times of the above Splatter simulations and generated 100 datasets for each of the above three different dropout rates. Figures 2B, S3A-F, and Table S4 (EXCEL) were derived from these datasets.

Real scRNA-seq datasets

Human brain scRNA-seq data. We used scRNA-seq data of 466 cells capturing the cellular complexity of the adult and fetal human brain at a whole transcriptome level [33]. Tag tables were downloaded from the data repository NCBI Gene Expression Omnibus (GEO access number: GSE67835) and combined into one table with columns representing cells and rows representing genes. We excluded the uncertain hybrid cells and remained 420 cells in eight cell types with the expression of 22085 genes. This dataset was used to generate Figures 2C-D and S3G, and Table S4.

Cell-cycle phase scRNA-seq data. To evaluate the performance of different imputation methods on identifying different cellular states of the same cell type, we analyzed a single-cell RNA-seq data from mESCs [34]. A set of 96 asynchronously dividing cells for each cell-cycle phase of G1, S, and G2M was captured using the Fluidigm C1 system, and sequencing libraries were prepared and processed. In this dataset, 288 mESCs were profiled and characterized by 38293 transcripts with a dropout rate of 74.4%. This dataset was used to generate Figures 3A-B and S3, and Table S6.

ERCC spike-in RNAs scRNA-seq data. In the above scRNA-seq dataset for mESCs, ERCC spike RNAs were added to each cell and sequenced. ERCC spike RNAs consist of 92 RNA transcripts in length of 250 to 2,000 nt, which are widely used in scRNA-seq experiments to remove the confounding noises from biological variability. Since RNA spike-in is added to samples with the identical amount to capture the technical noise, the readout for the spike-in RNAs should be free of cell-to-cell variability and the detected variance of expression, if exists, should only come from technical confounders other than biological contexts (e.g. cell types). Therefore, the expression profiles of spike-in RNAs that were added to individual cells should

not be able to cluster these cells into different subgroups regarding cell types or other biological states. Therefore, We used the ERCC spike-in read counts from the real scRNA-seq data for mESCs [34] to evaluate the imputation methods on denoising the technical variation without introducing extra noise. This data was used to generate Figures 2E and S3H-I, and Table S5.

Mouse ESCs scRNA-seq dataset for cell-cycle dynamics. 6885 mouse embryonic stem cells (mESC) were profiled using the droplet-microfluidic scRNA-seq approach with 1 biological replicate (933 cells) and 2 technical replicates (2509 and 3443 cells for each). The processed count matrix was downloaded from Gene Expression Omnibus (GEO) with the access ID GSE65525. All other nine imputation methods and scIGANs were used to impute the raw matrix with an exception that SCRABBLE and DrImpute failed to impute this data because take longer than a month to finish the imputation. This data was used to generate Figures 3C and S4C-L.

Cell cycle dynamics assessment was performed according to Figure 6E-F of [14]. Briefly, the Pearson's correlation was applied among a list of previously categorized 44 cell-cycle genes based on their expression across these 6.8k cells. Genes were ordered by hierarchical clustering on the correlation matrix and their previously categorized cell-cycle phases were indicated as linked dots representing cell-cycle oscillations (Figures 3C and S4C-L). Clustering measurements were also applied to the gene clusters against their pre-assigned cell-cycle phased (bar plots in Figures 3C and S4C-L), which represent the performances of imputation methods on clustering the genes across cells.

Human ESC scRNA-seq dataset for differential expression analysis. To compare the performance of different imputation methods on the identification of differentially expressed genes (DEGs), we utilize a real dataset with both bulk and single-cell RNA-seq experiments on human embryonic stem cells (ESC) and the differentiated definitive endoderm cells (DEC) [41]. This dataset includes six samples of bulk RNA-seq (four for H1 ESC and two for DEC) and scRNA-seq of 350 single cells (212 for H1 ESC and 138 for DEC). The percentage of zero

expression is 14.8% in bulk data and 49.1% in single-cell data. This dataset was used to generate Figures 4 and S5-S7.

We use scIGANs and nine other imputation methods to impute the gene expression for single cells and then use DESeq2 [42] to perform differential expression analysis on the raw and 10 imputed data, respectively. DEGs are genes with the absolute log fold changes (H1/DEC) ≥ 1.5, adjust-p ≤ 0.05, and base mean ≥ 10 (Figure 4A). A set of top 1000 DEGs (500 best up-regulated and 500 best down-regulated genes based on their adjust-p values) from bulk RNA-seq data were used to evaluate the correspondence between scRNA-seq and bulk RNA-seq data (Figures 4B and S5). To further evaluate the improvement of imputation on DEG identification, five signature genes highlighted in Figure 1c of the source paper [42] for H1 and DEC, respectively, were plotted out (Figures 4C and S6). The expression of two marker genes (SOX2 for H1 cell and CXCR4 for DEC cell) were overlaid to the UMAP space of single cells to show the expression signature of these two types of cells (Figures 4D-E and S7).

Time-course scRNA-seq data for cellular trajectory analysis. We utilize a time-course scRNA-seq data derived from the differentiation from H1 ESC to definitive endoderm cells (DEC) [41]. A total of 758 cells were profiled at 0 (n=92), 12 (n=102), 24 (n=66), 36 (n=172), 72 (n=138), and 96 (n=188) hours after inducing the differentiation from H1 ESCs to DECs (Figure 5A). We apply scIGANs and all other nine imputation methods to the raw scRNA-seq data with known time points and then reconstruct the trajectories.

Subsampling for robustness analysis. We subsampled the scRNA-seq data derived from human embryonic stem cells (ESC) and the differentiated definitive endoderm cells (DEC) [40]. This dataset has expression profiles of 350 single cells (212 for H1 ESC and 138 for DEC) across 19097 genes. Three different sampling strategies were used to generate different sub-datasets for robustness tests. These datasets were used to generate Figures 6 and S9.

1) datasets with a subset of genes that have top- and lower-mean of expressions across all 350 cells, denoted as mean.top and mean.low. Specifically, the expression matrix (genes in rows and cells in columns) was sorted by the row means (descending) and the first and last 5000 genes were selected, representing two subsets with high and low expressions, respectively. Then 1024 (32*32) genes were randomly picked from these 5000 genes to generate the two test datasets, mean.top and mean.low (Figures 6 and S9). These two datasets have the zero-rate of 6.34% (mean.top) and 97.25% (mean.low).

2) datasets with a subset of genes that have top- and lower-standard deviation (sd) of expressions across all 350 cells, denoted as sd.top and sd.low. Specifically, the expression matrix (genes in rows and cells in columns) was sorted by the row sd (descending) and the first and last 5000 genes were selected, representing two subsets with high and low expression standard deviations, respectively. Then 1024 (32*32) genes were randomly picked from these 5000 genes to generate the two test datasets, sd.top and sd.low (Figures 6 and S9). These two datasets have the zero-rate of 8.72% (mean.top) and 92.42% (mean.low).

3) dataset with a subset of 1024 genes randomly selected from all 19097 genes, denoted as global.random. It has the zero-rate of 49.51%.

Implementation and availability

scIGANs is implemented in Python 3.6 and R 3.6.1 with an interface wrapper script. An expression matrix of the single cells is the only required input file. Optionally, a file including the cell labels (cell type or subpopulation information) can be provided to direct scIGANs for cell type-specific imputation. If there are no prior cell labels provided, scIGANs will pre-cluster the cells using a spectral clustering method. ScIGANs can run on either CPUs or GPUs. The output is the imputed expression matrix of the same dimensions, of which only the true zero values will be imputed without change other expression values. The whole package with a usage tutorial is available at GitHub (https://github.com/xuyungang/scIGANs).

Availability of data and codes for reproducibility

The original sources and preprocesses of all data are described in Methods. The processed datasets and codes used to reproduce the Figures and Tables are available at GitHub (https://github.com/xuyungang/scIGANs_Reproducibility).

Statistical information

All statistical tests are implemented by R (version 3.6.1). Specifically, the Pearson correlation tests (Figures 4B and S5) were done by cor.test() with default parameters; the student's t-tests (Figures 6C-D and S9) were done by t.test() with default parameters; the differentially expressed genes (DEGs) were identified by DESeq2 with the p-adjust <=0.05, log2FoldChange >= 1.5, and baseMean >= 10 (Figures 4A-B and S5).

Qantitative measurments of single cell clusters

We use 11 numeric metrics to quantitate the clustering of single cells. RI, the Rand index, is a measure of the similarity between two data clusterings. ARI, the adjusted Rand index, is adjusted for the chance grouping of elements. MI, mutual information, is used in determining the similarity of two different clusterings of a dataset. As such, it provides some advantages over the traditional Rand index. AMI, adjusted mutal information, is a variation of mutual information used for comparing clusterings. VI, variation of information, is a measure of the distance between two clusterings and a simple linear expression involving the mutual information. NVI the normalized VI. ID and NID refer to the information distance and normalized information distance. All these metrics are computed using clustComp() from R package 'aricode' (https://cran.r-project.org/web/packages/aricode/). F score (also F1-score or F-measure) is the harmonic mean of precision and recall. AUC, area under the receiver operating characteristic (ROC) curve, is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. ACC, accuracy. The above three classification metrics are defined by compare the independent clustering of cells to the true cell lables.

Clustering was done using prediction() from the R package SC3 [50]. The in-house R scripts for these metrics are provided in the codes for reproducibility.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's Contributions

YX, ZZ, and XZ conceived the study. ZZ developed the scIGANs model and YX wrapped it up to a package. YX analyzed all scRNA-seq datasets, interpreted the results and wrapped up the reproducibility codes on GitHub (https://github.com/xuyungang/scIGANs_Reproducibility). ZZ, LY, JL, and ZF helped to test the method and reproduce the analyses. YX wrote the manuscript and all authors revised it. All authors read and approved the final version of the manuscript.

Materials and Correspondence

The correspondence and material request should be addressed to Yungang Xu (yungang.xu@uth.tmc.edu).

Acknowledgments

This work was funded by the National Institutes of Health (NIH) [R01CA241930, R01GM123037 and AR069395].

Additional files

Additional file 1 (PDF): Supplementary Figures S1-S9.

Additional file 2 (PDF): Supplementary Tables S1, S2, and Tables S4-S6.

Additional file 3 (XLSX): Supplementary Table S3.

References

1. van Dijk, D., et al., *Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.* Cell, 2018. **174**(3): p. 716-729 e27.
2. van Dijk, D., et al., *Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.* Cell, 2018. **174**(3).
3. Li, W. and J. Li, *An accurate and robust imputation method scImpute for single-cell RNA-seq data.* Nature Communications, 2018. **9**(1).
4. Huang, M., et al., *SAVER: gene expression recovery for single-cell RNA sequencing.* Nature Methods, 2018. **15**(7): p. 539-542.
5. Gong, W., et al., *DrImpute: imputing dropout events in single cell RNA sequencing data.* BMC Bioinformatics, 2018. **19**(1): p. 220.
6. Chen, M. and X. Zhou, *VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies.* Genome Biology, 2018. **19**(1): p. 196.
7. Eraslan, G., et al., *Single-cell RNA-seq denoising using a deep count autoencoder.* Nature Communications, 2019. **10**(1): p. 390.
8. Wagner, F., D. Barkley, and I. Yanai, *Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis.* bioRxiv, 2019: p. 655365.
9. Arisdakessian, C., et al., *DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data.* Genome Biol, 2019. **20**(1): p. 211.
10. Peng, T., et al., *SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data.* Genome Biology, 2019. **20**(1): p. 88.
11. Mattei, P.A. and F.-J. on Machine, *MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets.* International Conference on Machine …, 2019.
12. Zhang, H., P. Xie, and X.-E. preprint arXiv, *Missing value imputation based on deep generative models.* arXiv preprint arXiv:1808.01684, 2018.
13. Mattei, P.A. and F.-J. preprint arXiv, *missiwae: Deep generative modelling and imputation of incomplete data.* arXiv preprint arXiv:1812.02633, 2018.
14. Lopez, R., et al., *Deep generative modeling for single-cell transcriptomics.* Nature Methods, 2018. **15**(12): p. 1053-1058.
15. Goodfellow, I., J. Pouget-Abadie, and M.M. in neural …, *Generative adversarial nets.* Generative adversarial nets, 2014.
16. Radford, A., L. Metz, and C.S. preprint arXiv, *Unsupervised representation learning with deep convolutional generative adversarial networks.* Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.
17. Chen, X., et al., *Infogan: Interpretable representation learning by information maximizing generative adversarial nets.* Advances in neural …, 2016.
18. Miyato, T., et al., *Spectral normalization for generative adversarial networks.* arXiv preprint arXiv …, 2018.
19. Ghahramani, A., F.M. Watt, and N.M. Luscombe, *Generative adversarial networks simulate gene expression and predict perturbations in single cells.* bioRxiv, 2018: p. 262501.
20. Ahmed, F., M. Arjovsky, and D.-V. in neural …, *Improved training of wasserstein gans.* Advances in neural …, 2017.
21. Yoon, J., J. Jordon, and M.J.a.p.a. van der Schaar, *GAIN: Missing Data Imputation using Generative Adversarial Nets.* 2018.
22. Ledig, C., et al. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*. in *CVPR*. 2017.

23. Brock, A., et al., *Neural photo editing with introspective adversarial networks.* 2016.
24. Wolterink, J.M., et al., *Generative adversarial networks for noise reduction in low-dose CT.* 2017. **36**(12): p. 2536-2545.
25. Zhang, H., V. Sindagi, and V.M.J.a.p.a. Patel, *Image de-raining using a conditional generative adversarial network.* 2017.
26. Chen, Q. and V. Koltun. *Photographic image synthesis with cascaded refinement networks*. in *IEEE International Conference on Computer Vision (ICCV)*. 2017.
27. Marouf, M., et al., *Realistic in silico generation and augmentation of single cell RNA-seq data using Generative Adversarial Neural Networks.* bioRxiv, 2018: p. 390153.
28. Marouf, M., et al., *Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks.* Nature Communications, 2020. **11**(1): p. 166.
29. Lin, P., M. Troup, and J.W.K. Ho, *CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data.* Genome Biology, 2017. **18**(1).
30. Zare, H., et al., *Data reduction for spectral clustering to analyze high throughput flow cytometry data.* BMC Bioinformatics, 2010. **11**: p. 403.
31. Zappia, L., B. Phipson, and A. Oshlack, *Splatter: simulation of single-cell RNA sequencing data.* Genome Biol, 2017. **18**(1): p. 174.
32. Peng, T., et al., *SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data.* Genome Biol, 2019. **20**(1): p. 88.
33. Spyros, D., et al., *A survey of human brain transcriptome diversity at the single cell level.* Proceedings of the National Academy of Sciences, 2015. **112**(23): p. 7285-7290.
34. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nature Biotechnology, 2015. **33**(2): p. 155-160.
35. Paul, F., et al., *Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors.* Cell, 2015. **163**(7): p. 1663-1677.
36. Wilson, N.K., et al., *Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations.* Cell stem cell, 2015. **16**(6): p. 712-724.
37. Buettner, F., et al., *Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells.* Nature Biotechnology, 2015. **33**(2): p. 155-160.
38. McDavid, A., G. Finak, and R. Gottardo, *The contribution of cell cycle to heterogeneity in single-cell RNA-seq data.* Nature Biotechnology, 2016. **34**(6): p. 591.
39. Rapsomaniki, M., et al., *CellCycleTRACER accounts for cell cycle and volume in mass cytometry data.* Nature Communications, 2018. **9**(1): p. 632.
40. Klein, A.M., et al., *Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells.* Cell, 2015. **161**(5): p. 1187-1201.
41. Chu, L.F., et al., *Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm.* Genome Biol, 2016. **17**(1): p. 173.
42. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
43. Chen, H., et al., *Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM.* Nature Communications, 2019. **10**(1): p. 1903.
44. Bendall, S.C., et al., *Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development.* Cell, 2014. **157**(3): p. 714-725.
45. Qiu, X., et al., *Reversed graph embedding resolves complex single-cell trajectories.* Nature Methods, 2017. **14**(10): p. 979-982.

46. Schiebinger, G., et al., *Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming.* Cell, 2019. **176**(Cell Tissue Res. 331 2008): p. 928.
47. Yang, Q., et al., *Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss.* IEEE Trans Med Imaging, 2018. **37**(6): p. 1348-1357.
48. Gulrajani, I., et al., *Improved Training of Wasserstein GANs.* arXiv, 2017.
49. Berthelot, D., T. Schumm, and L. Metz, *BEGAN: Boundary Equilibrium Generative Adversarial Networks.* BEGAN: Boundary Equilibrium Generative Adversarial Networks, 2017.
50. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.

Figure 1. Overview of generative adversarial networks for single-cell RNA-seq imputation (scIGANs). The expression profile of each cell is reshaped to a square image, which is fed to the GANs (Supplementary Figure S1A). The trained generator is used to generate a set of realistic cells, of which the k-nearest neighbors (KNN) are used to impute the raw scRNA-seq expression matrix (Supplementary Figure S1B). MSE, mean squared error.

Also see Figures S1.

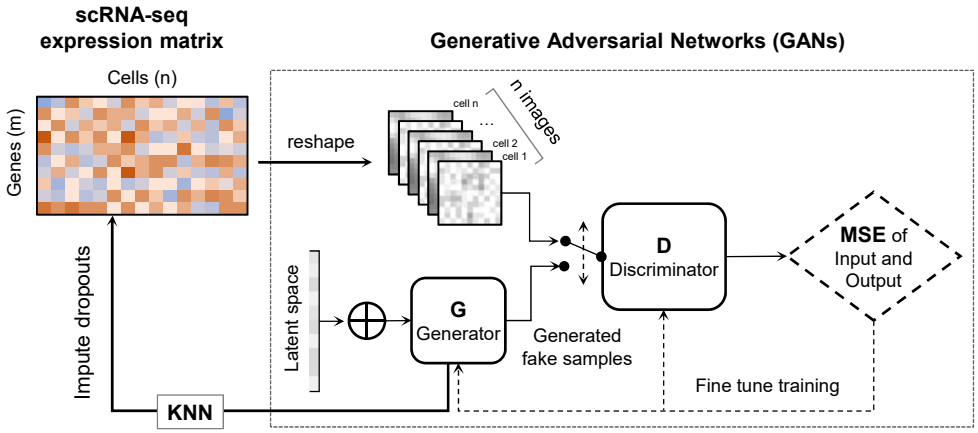


Figure 2. ScIGANs recovers single-cell gene expression from dropouts without extra noise. A. The UMAP plots of the CIDR simulated scRNA-seq data for Full, Dropout, and imputed matrix by 10 methods. Multiple clustering measurements are provided in Supplementary Figure S2A and Table S1. **B.** The adjusted rand index (ARI), a representative clustering measurement to indicate performance and robustness of all methods on the Splatter simulated data with three different dropout rates (71%, 83%, 87%) and 100 replicates for each. The plots of other selected measurements are provided in Supplementary Figures S3A-F and the full list of clustering measurements provided in Supplementary Table S3. **C.** The selected UMAP plots of real scRNA-seq data for human Brain; the plots of all other imputation methods are provided in Supplementary Figures S3G. **D.** The selected clustering measurements for scRNA-seq data of human Brian. AUC, area under the ROC curve; ARI, adjusted rand index; F score, the harmonic mean of precision and recall; NMI, normalized mutual information. Full list of all considered clustering measurements are provided in Supplementary Table S4. **E.** The evaluation of robustness in avoiding extra noise using scRNA-seq data of spike-in RNAs. All UMAP plots are provided in Supplementary Figure S3H.

Also see Figures S2-S3.

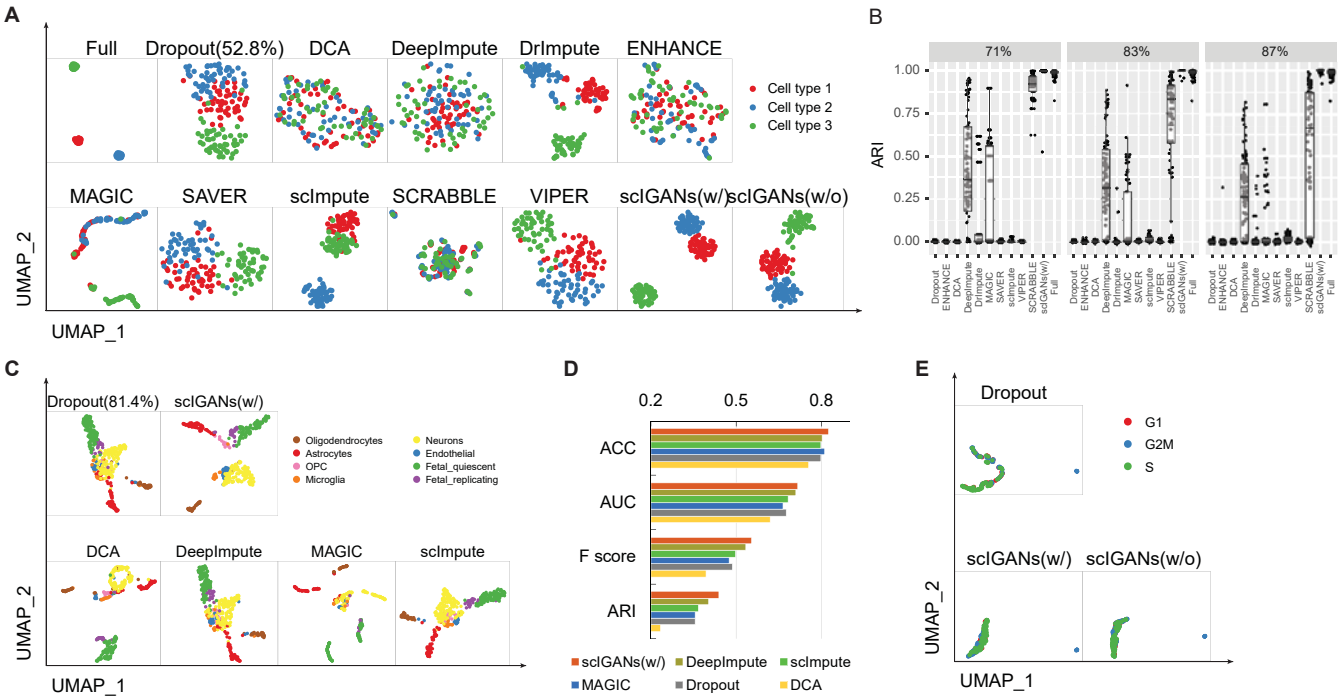


Figure 3. ScIGANs enables the identification of subcellular states. A. The UMAP plots of the real scRNA-seq data with known cell-cycle states. Full list of all considered clustering metrics are provided in Supplementary Figure S4A and Table S6. **B.** Cells are projected to the cell-cycle phase spaces based on collections of cell-cycle genes. The plots for all other methods are provided in Supplementary Figure S4B. **C.** Cell cycle dynamics shown as the hierarchical clustering of 44 cell-cycle-regulated genes across 6.8k mouse ESCs. Full dynamic cell-cycle profiles from before and after imputation by different methods are provided in Supplementary Figure S4C-L. The bar charts show the quantitative concordance between the assigned cell-cycle phases by hierarchical clustering and the true phases for which these genes serve as markers. F score, the harmonic mean of precision and recall; AUC, area under the ROC curve; ACC, accuracy.

Also see Figure S4.

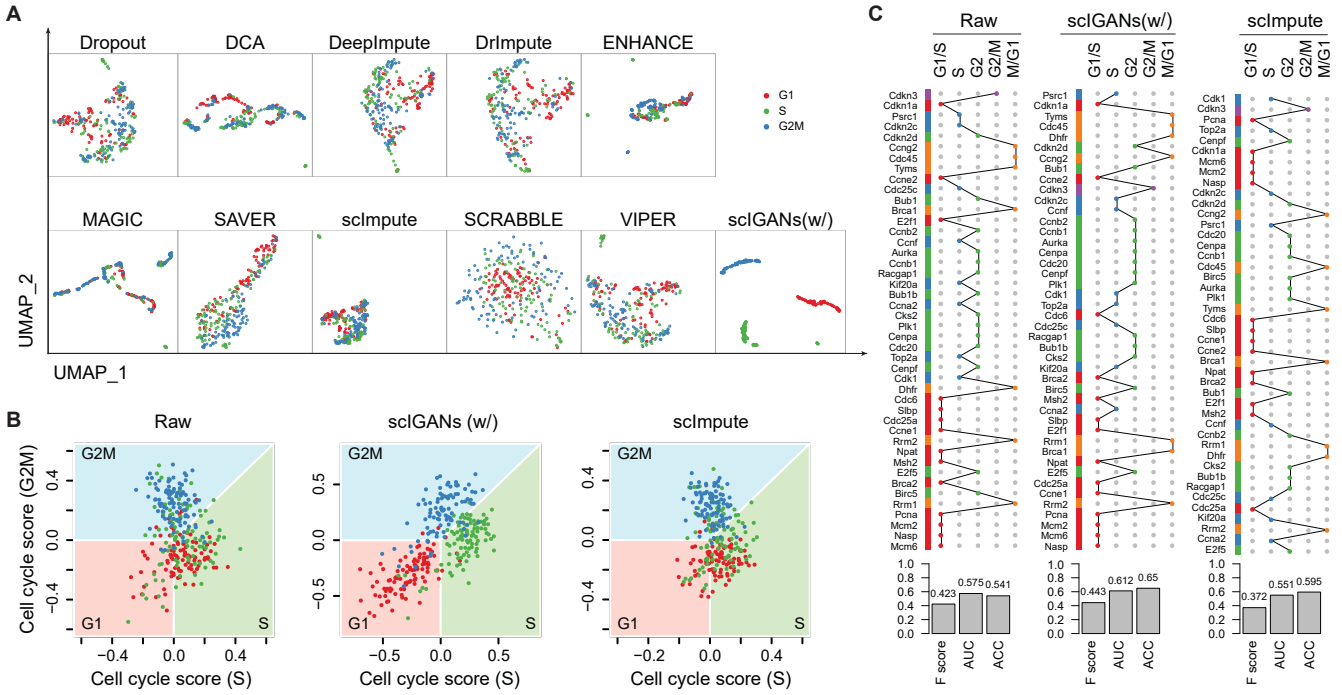


Figure 4. ScIGANs increases the correspondence of differential expression between single-cell and bulk RNA-seq. A. The correspondence of differentially expressed genes (DEGs) between bulk and single-cell RNA-seq with different imputation approaches. **B.** The correlations between log fold changes of differentially expressed genes from bulk and single-cell RNA-seq. Detailed legends and the plots of all considered imputation methods are provided in Supplementary Figure S5. **C.** The expression for one of five selected signature genes of H1 and DEC cells, respectively. All plots of other genes with different imputation methods are provided in Supplementary Figure S6. **D-E.** The UMAP plots of the single cells overlaid by the expression of SOX2 and CECR4, which is the marker of H1 and DEC, respectively. Raw (D) and scIGANs imputed (E) matrix are shown and all other methods are provided in Supplementary Figure S7.

Also see Figure S5-S7.

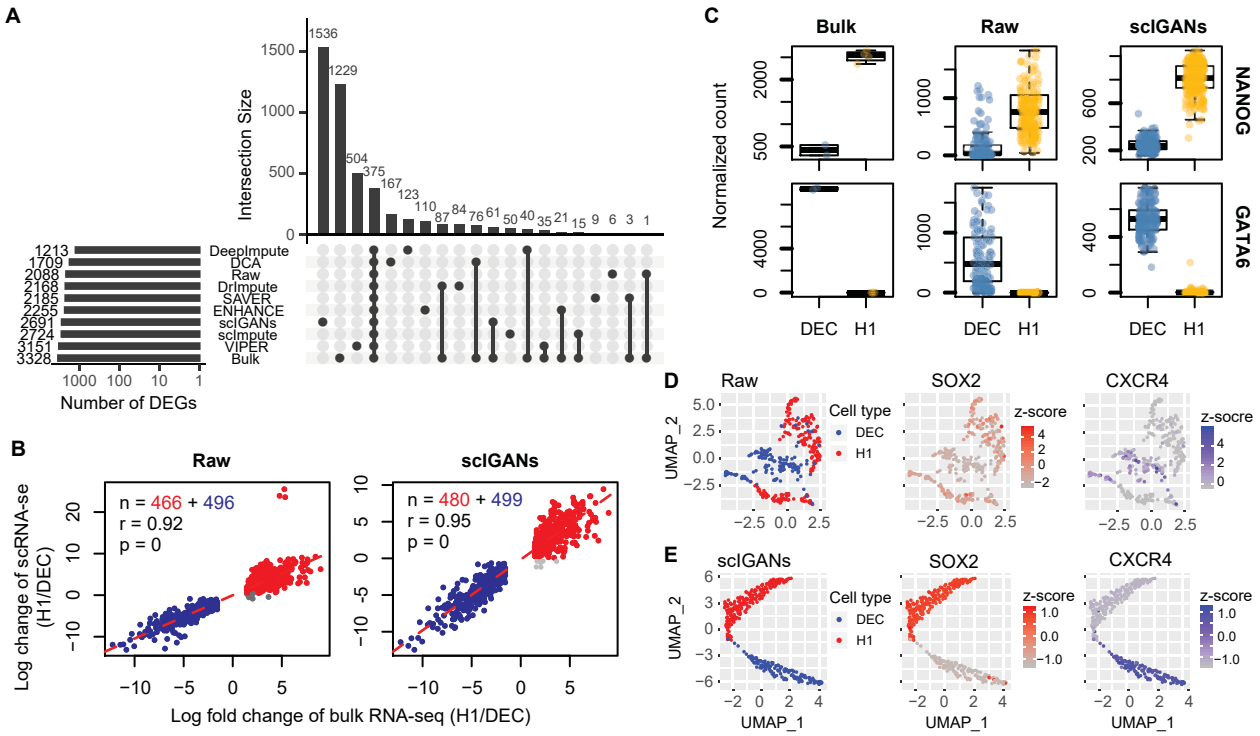


Figure 5. ScIGANs improves time course analysis and reconstruction of cellular trajectory from scRNA-seq data. A. The time points of scRNA-seq sampling along the differentiation from pluripotent state (H1 cells) through mesendoderm to definitive endoderm cells (DEC). **B-C.** The trajectories reconstructed by monocle3 from the raw (C) and scIGANs imputed (D) scRNA-seq data. **D-E.** The expression of two pluripotent (left) and DEC (right) signature genes are shown in the order of the pseudotime. The plots of all other imputation methods are provided in Supplementary Figure S8.

Also see Figure S8.

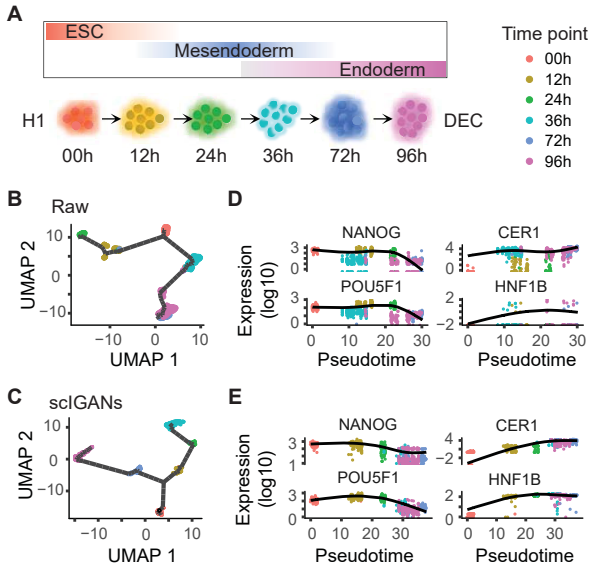
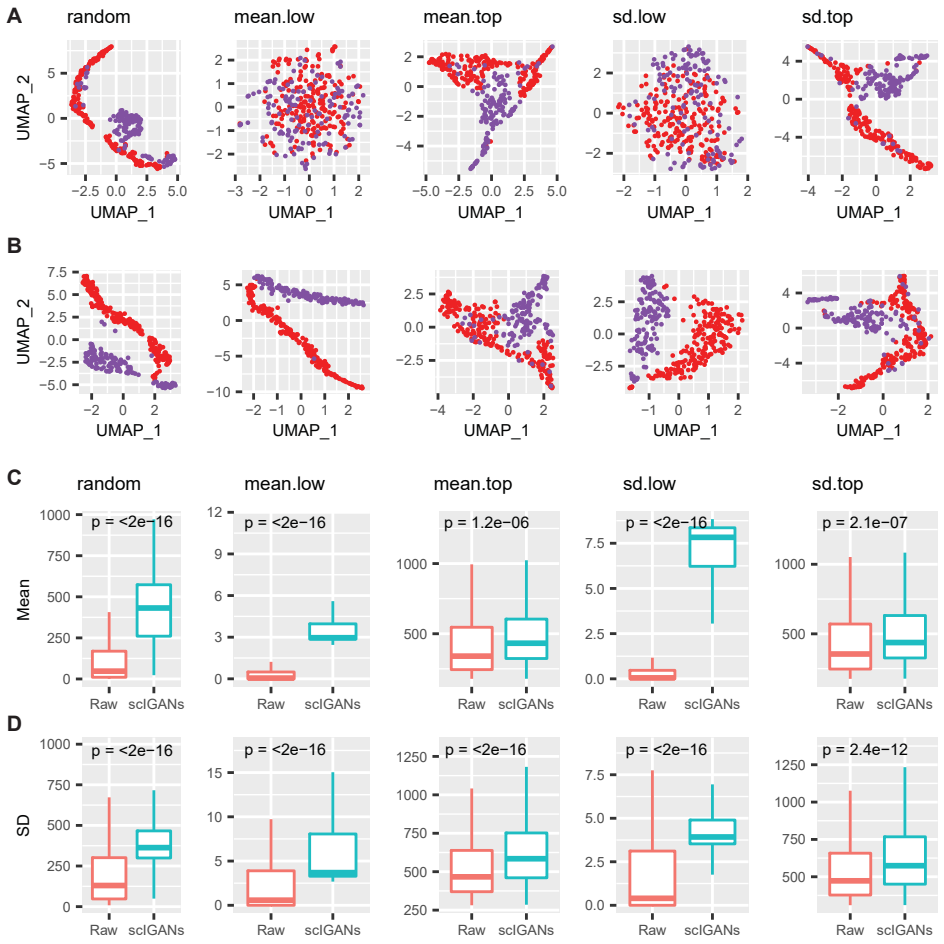


Figure 6. ScIGANs is robust to small set of genes with very low expression or cell-to-cell variance. A-B. The UMAP visualizations of H1 and DEC cells using only 1024 genes from raw (A) or scIGANs imputed (B) expression matrix based on three different sampling strategies. The sampling strategies are described in Methods. **C-D.** The boxplots show the mean (C) or standard deviation (sd, D) of the 1024 sampled genes before and after scIGANs imputation; p, the p-value of the Student's t-test (two-side). The same series of plots for all other imputation methods are provided in Supplementary Figure S9.

Also see Figure S9.