Pathway-Based Single-Cell RNA-Seq Classification, Clustering, and Construction of Gene-Gene Interactions Networks Using Random Forests

Hailun Wang, *Student Member, IEEE*, Pak Sham, Tiejun Tong, and Herbert Pang

***Abstract*—Single-cell RNA-Sequencing (scRNA-Seq), an advanced sequencing technique, enables biomedical researchers to characterize cell-specific gene expression profiles. Although studies have adapted machine learning algorithms to cluster different cell populations for scRNA-Seq data, few existing methods have utilized machine learning techniques to investigate functional pathways in classifying heterogeneous cell populations. As genes often work interactively at the pathway level, studying the cellular heterogeneity based on pathways can facilitate the interpretation of biological functions of different cell populations. In this paper, we propose a pathway-based analytic framework using Random Forests (RF) to identify discriminative functional pathways related to cellular heterogeneity as well as to cluster cell populations for scRNA-Seq data. We further propose a novel method to construct gene-gene interactions (GGIs) networks using RF that illustrates important GGIs in differentiating cell populations. The co-occurrence of genes in different discriminative pathways and 'cross-talk' genes connecting those pathways are also illustrated in our networks. Our novel pathway-based framework clusters cell populations, prioritizes important pathways, highlights GGIs and pivotal genes bridging cross-talked pathways, and groups co-functional genes in networks. These features allow biomedical researchers to better understand the functional heterogeneity of different cell populations and to pinpoint important genes driving heterogeneous cellular functions.**

***Index Terms*—cellular heterogeneity, gene-gene interactions (GGIs) networks, functional pathways, Random Forests (RF), single-cell RNA-Sequencing.**

Manuscript received March 1, 2019; revised June 17, 2019 and September 2, 2019; accepted September 18, 2019. Date of publication October 1, 2019; date of current version June 5, 2020. This work has been partially supported by the Research Grants Council - General Research Fund no. 17157416. *(Corresponding author: Herbert Pang.)*

H. Wang and H. Pang are with the School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China (e-mail: u3005316@connect.hku.hk; herbpang@hku.hk).

P. Sham is with the Department of Psychiatry and Centre for Genomic Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China (e-mail: pcsham@hku.hk).

T. Tong is with the Department of Mathematics, Hong Kong Baptist University, Hong Kong SAR, China (e-mail: tongt@hkbu.edu.hk).

Digital Object Identifier 10.1109/JBHI.2019.2944865

I. Introduction

SINGLE-CELL RNA-Sequencing (scRNA-Seq) is one of the latest advances in high-throughput sequencing technologies. It enables the resolution of gene expression profiling at the level of individual cells and allows to study biological processes with a smaller number of cells compared to the traditional bulk RNA-Sequencing (RNA-Seq) [1]. scRNA-Seq also provides insights into cell-specific questions including: (a) cell lineage and differentiation states [2], [3], (b) identification of cell types or sub-cell types [4], [5], and (c) intra-tumor heterogeneity [6], [7].

The rapid growth of interest in scRNA-Seq has led to an increasing demand of appropriate computational methods for understanding the cellular heterogeneity. A number of machine learning based methods have been developed for unsupervised clustering of cell populations for scRNA-Seq data. Related methods include clustering cells based on reduced dimensionality of data through principal component analysis (PCA) [8], t-distributed stochastic neighbor embedding (t-SNE) [9], and diffusion maps [10]; multiple kernels learning [11]; consensus clustering [12]; multiobjective evolutionary clustering imposed with non-negative matrix factorization [13]; and Random Forests based similarity learning [14], [15]. Despite these advancements, assigning biological functions to clustered or known cell populations from scRNA-Seq data is still a challenge [16]. As genes often work interactively but not individually, identifying discriminative pathways can improve our understanding of functional heterogeneity of cell populations. In addition, few existing machine learning based methods for scRNA-Seq data have been adapted to study cellular heterogeneity at the pathway level. Gene set enrichment analysis (GSEA) developed for bulk RNA-Seq can be applied to scRNA-Seq. However, as scRNA-Seq data is sparser [17] and noisier than the bulk RNA-Seq data [18], conventional methods for bulk RNA-Seq data may not be optimal for scRNA-Seq data.

Inspired by the biological importance of pathways and the lack of pathway-relevant analytic methods for scRNA-Seq, we developed a pathway-based framework for analyzing scRNA-Seq using Random Forests (RF). RF is an ensemble learning method based on classification trees [19]. Superiority of RF over other popular learning methods for bulk RNA-Seq gene expression data has been demonstrated [20]–[22]. The idea of prioritizing pathways through RF was previously proposed for bulk RNA-Seq data for the binary classification [23]. In this paper, we extend the idea to the field of scRNA-Seq data for identifying discriminative functional pathways in explaining the cellular heterogeneity among multiple cell populations. Our new framework can cluster cell populations by utilizing pathway information, and can construct gene-gene interactions (GGIs) networks by connecting 'cross-talk' genes of discriminative pathways using RF. We have also implemented our framework in a publicly available R package *scPathwayRF* (single-cell pathway-based random forests classification, clustering and construction of GGIs networks), which can be accessed from the following link: http://web.hku.hk/∼herbpang/scPathwayRF.html.

II. Background

RF is the backbone machine learning algorithm for our novel classification, clustering, and construction of GGIs networks tool. RF is an ensemble classifier that aggregates multiple independent decision trees [19]. Each tree is built based on a bootstrapped set of observations with the same dimension as the original set of observations. Since samples are randomly selected with replacement, some samples are left out in each tree in RF, which are called out-of-bag (OOB) samples. Each set of OOB samples in a tree in RF can be used as a built-in testing set. During the construction of the tree, a random subset of input features is selected to split each node of the tree. Classification and regression tree (CART) methodology [24] is adapted to grow the tree to maximum depth without pruning. The prediction of the class for an observation is the majority vote over all trees in the RF classifier. In RF, randomness is introduced in both sampling and feature selection for node splitting. Consequently, this unique character of RF makes it robust to outliers and noise [19].

III. Materials and Methods

In this section, we describe the methodology of our proposed approach. In Section III-A, we present RF classification for prioritizing discriminative pathways. This is followed by the presentation of RF clustering for grouping cell populations in Section III-B. In Section III-C, we present the methodology for GGIs networks construction. Following the introduction of methodologies in Sections III-A to III-C, we present the methods used for the comparison of our proposed RF classification and RF clustering with other algorithms in Section III-D. We describe in Section III-E a simulation study for understanding the influence of informative genes on the performance of pathway-based RF for scRNA-Seq data. A schematic diagram of the whole proposed analytic framework is illustrated in Supplementary Fig. 1(a).[1]

Let $X \in \mathbb{R}^{m_e \times n}$ denote a scRNA-Seq gene expression matrix of size m_e-by-n, corresponding to the number of expressed genes by the number of observed cells, respectively. Suppose n observed cells in X have been labelled into different cell populations based on the known information or pre-clustering. Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ denote the labels of cell populations for n cells. Let $\mathrm{P} \in \mathbb{R}^{m_g \times p}$ denote a binary matrix of functional pathways database which contains m_g number of genes and p number of pathways, and let

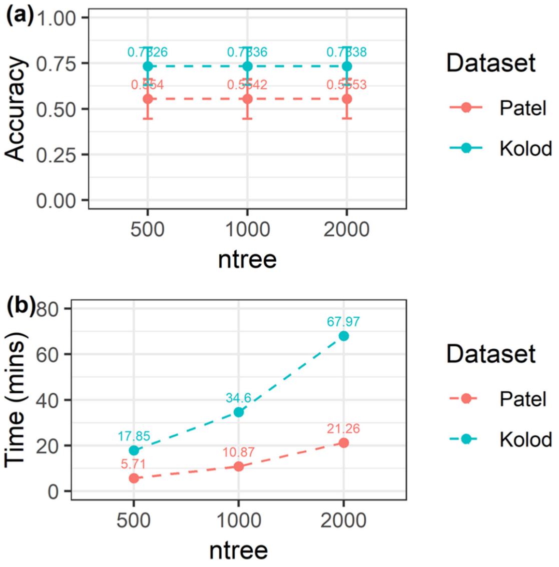

Fig. 1. Effect of ntree on the classification performance and speed. (a) The classification accuracy has little change as ntree increases from the default value (ntree = 500). (b) The computational time increases proportionally with ntree.

$$\mathrm{P}_{ij} = \begin{cases} 1 \text{ if the } i\text{th gene is involved in the } j\text{th pahtway,} \\ 0 \text{ otherwise,} \end{cases} \tag{1}$$

where $i = 1, \ldots, m_g$, and $j = 1, \ldots, p$. Two sources of pathways databases: BioCarta (http://www.biocarta.com/) and KEGG [25], have been added as the built-in choices of databases in *scPathwayRF* package. Given X, $\boldsymbol{y}$, and P, our framework enables the identification of discriminative pathways through classification of cell populations, clustering of cell populations, and construction of GGIs networks using pathway information and RF.

A. Discriminative Functional Pathways in Differentiating Cell Populations

Recall that the scRNA-Seq gene expression matrix X contains m_e unique genes and the functional pathways database matrix P contains m_g unique genes. Suppose M_e is the set of unique genes in X, and M_g is the set of unique genes in P. Let M_I denote the intersection $M_e \cap M_g$, which contains m_I number of genes with $m_I \leq \min(m_e, m_g)$. In the first step, a submatrix $X_I \in \mathbb{R}^{m_I \times n}$ of X, and a submatrix $P_I \in \mathbb{R}^{m_I \times p}$ of P are extracted and stored, where rows of both X_I and P_I are m_I.

Let $\mathbf{p}_j, j = 1, \ldots, p$, denote the j^{th} column vector of P_I, which represents the j^{th} pathway in P_I. Define the pathway size n_{size} as the number of expressed genes in the j^{th} pathway, which equals

[1] Supplementary materials can be downloaded from the link http://web.hku.hk/~herbpang/IEEE_JBHI_Supplementary2019Jun.pdf.

to the sum of the j^{th} column of P_I:

$$n_{size} = \sum_{a=1}^{m_I} \mathrm{P}_{Iaj}. \tag{2}$$

A user-defined cutoff n_{cutoff} (default $n_{cutoff} = 5$, for the rationale of the choice of 5, please refer to Supplementary Methods M1) is provided to filter out small pathways. For the pathway $\mathbf{p}_j$ satisfying $n_{size} \geq n_{cutoff}$, the corresponding observation set used by the RF classifier is obtained through an element-wise multiplication between the column vector of X_I and $\mathbf{p}_{j.}$ Unexpressed genes are trimmed. And an $n_{size} \times n$ matrix is generated for building a RF classifier for $\mathbf{p}_{j.}$

The goodness G of each investigated pathway in differentiating cell populations is estimated by averaging classification accuracy of k-fold (default $\mathrm{k} = 5$) cross validation (CV) of a pathway-based classifier. Indexes of sampling of k-fold CV are kept consistent across all investigated pathways to prevent the potential sampling bias. Suppose $\hat{\boldsymbol{y}}_f = \{\hat{y}_{f_1}, \ldots, \hat{y}_{f_n}\}$, $f \in \{1, \ldots, \mathrm{k}\}$ is the set of predicted cell populations for n cells in the fth fold classification. Then the goodness G can be calculated as

$$G_f = \frac{\sum_{x=1}^{n} 1\left(\hat{y}_{f_x} = y_x\right)}{n}, \tag{3}$$

$$G = \frac{1}{\mathrm{k}} \sum_{f \in \{1, \cdots, \mathrm{k}\}} G_f, \tag{4}$$

where 1 is an indicator function, and G_f represents the classification accuracy in the f^{th} fold CV. All investigated pathways are prioritized based on their respective G values. By default, pathways with the top 10 highest G are selected as top discriminative functional pathways in differentiating cell populations.

B. Pathway-Based Random Forests Clustering

RF has been applied previously in an unsupervised manner to learn the similarity of clustering samples by using both the original data and a newly generated synthetic data from the original data [14], [26], and [27], and temporary classes defined by discriminative features [15]. Here we propose an approach using RF for clustering scRNA-Seq data based on gene expression levels and pathways. A set of intermediate cluster labels is generated by performing partitioning around medoids (PAM) based on a similarity matrix S_{gene}. S_{gene} is calculated by averaging RF learned similarity matrices from dimensionally reduced matrices of three gene expression matrices X, X_I (see Section III-A), and X_u (a matrix of genes unannotated to any pathway). Relevant pathways are identified via classification using the intermediate cluster labels. Two popular dimension reduction methods PCA [28] and t-SNE [29] are applied. R package *Rtsne* [30] is used to perform t-SNE. When the number of clusters is unknown, the function pamk in R package *fpc* [31] will estimate the number of clusters by the average silhouette width to perform PAM. Otherwise, PAM will be based on users' input on the number of clusters. The set of intermediate cluster labels is used as class labels to build pathway-based RF. The elbow point [32] from sorted OOB errors of pathway-based RF, or a user-defined number, is used as the cutoff to select relevant pathways for clustering. We define a similarity matrix based on pathways, $\mathrm{S}_{pathway}$, by averaging similarity matrices learned from RF built on relevant pathways. A final similarity matrix S for cells is calculated by averaging S_{gene} and $\mathrm{S}_{pathway}$. The final cluster labels are determined by performing PAM based on the final similarity matrix S. Supplementary Fig. 1(b) visualizes the workflows of the proposed clustering method.

C. GGIs Networks Using Random Forests

After discriminative pathways are selected, important genes involved in these pathways are then chosen to construct the GGIs networks. There are two types of measures of variable importance offered by RF: (a) mean decrease in accuracy (MDA), and (b) Gini impurity index [26]. We use MDA as the measure to select important genes since the Gini measure is not as reliable as MDA [26]. The calculation of MDA is based on the sample margins. The margin of a sample is defined as the proportion of vote for true class minus the maximum proportion of vote for other classes in trees [19].

The MDA of a feature is defined as the average lowering of the margin across samples when this feature is permuted [26]. The detailed calculation of the margin and MDA of RF are shown in Supplementary Methods M2. Genes involved in each discriminative pathway are ranked by their MDAs, and by default, the top 10% of them with positive MDAs are considered as important genes (see Supplementary Methods M1 for the rationale of the default 10% threshold).

Suppose a total of n_{imp} important genes are selected. We use RF to find the potential interaction between two important genes in correctly predicting cell populations. Consider the set $\mathbf{T} = \{\mathrm{T}_1, \ldots, \mathrm{T}_\mathrm{b}\}$ of all trees in the RF classifier built on important genes as features. For each tree T_b, we trace back the decision paths that correctly predict the labels of the OOB samples in this tree. An interaction of two genes is defined as the occurrence that both genes are used in the same decision path in the tree $\mathrm{T}_{\mathrm{b}.}$ A matrix $\mathrm{I}^{\mathrm{T}_\mathrm{b}} \in \mathbb{R}^{n_{imp} \times n_{imp}}$ indicating GGIs based on the tree T_b can be defined as

$$\mathrm{I}_{uv}^{\mathrm{T}_\mathrm{b}} = \begin{cases} 1 \text{ if the } u\text{th gene interacts with the } v\text{th } gene, \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

Based on the definition, an interaction only considers two different genes simultaneously used in the decision for the correct prediction, thus $\mathrm{I}_{uv}^{\mathrm{T}_\mathrm{b}} = 0$ when $u = v$. And let $\mathrm{I} \in \mathbb{R}^{n_{imp} \times n_{imp}}$ denote a matrix representing the prevalence of GGIs among trees in the forest, which can be calculated as

$$\mathrm{I} = \frac{1}{|\mathbf{T}|} \sum_{\mathrm{T}_\mathrm{b} \in \mathbf{T}} \mathrm{I}^{\mathrm{T}_\mathrm{b}}, \tag{6}$$

where $|\mathbf{T}|$ is the length of set $\mathbf{T}$, i.e., the total number of trees in the forest.

A GGIs network for important genes is then constructed. Each important gene is a node in the network, and in total there are n_{imp} number of nodes. The undirected edge between two nodes indicates the prevalence of the interaction between two genes. To make the graph of network visual friendly, we allow users to calibrate the number of undirected edges using graph density D of the network, which is defined as

$$\mathrm{D} = \frac{2\mathrm{E}}{n_{imp} \times (n_{imp} - 1)}, \tag{7}$$

where E is the number of edges in the graph. The default graph density is 0.05, which can be adjusted based on the desired visual effect. Let $w(u, v)$ denote the weight of the edge between the u^{th} node and the v^{th} node. We define the weight to equal the prevalence of the RF-based interaction between the u^{th} gene and the v^{th} gene, i.e.,

$$w(u, v) = \mathrm{I}_{uv} = \mathrm{I}_{vu}. \tag{8}$$

Edges denoting the top prevalent interactions are drawn according to the desired graph density. We further use a co-occurrence matrix $\mathrm{C} \in \mathbb{R}^{n_{imp} \times n_{imp}}$ to measure the number of common discriminative pathways that genes sharing as the layout of the network. Genes sharing a higher number of common pathways are located more closely in the graph of the network. We extract a submatrix $\mathrm{P}_{imp} \in \mathbb{R}^{n_{imp} \times 10}$ from the pathway database matrix P_I, where rows in P_{imp} are important genes and columns in P_{imp} are discriminative functional pathways. The co-occurrence matrix C is defined as

$$\mathrm{C} = \mathrm{P}_{imp} \mathrm{P}_{imp}^{\mathrm{T}}, \tag{9}$$

where $\mathrm{P}_{imp}^{\mathrm{T}}$ is the transpose of P_{imp}. The diagonal value of C_{rr} is exactly equal to the total number of discriminative pathways the r^{th} gene involved in. In our network, the node size and the node color are scaled to be proportional to the value of C_{rr} corresponding to the r^{th} gene. 'Cross-talk' genes linking multiple number of discriminative pathways are then highlighted. We also provide an assessment of the influence of 'cross-talk' genes on the cellular heterogeneity using RF (see Supplementary Methods M3 for details).

R package *randomForest* version 4.6-14 [33] is used to construct a RF classifier. R package *iRF* version 2.0.0 [34] is used to decipher decision paths of trees in a RF classifier. A function to plot the graph of our proposed network is implemented in our package by employing R package *igraph* version 1.2.2 [35].

D. Comparison With Other Classification and Clustering Algorithms for Cell Populations

We compare the performance in classifying cell populations of RF with other typical classifiers as well as a deep learning algorithm. The chosen algorithms include: (a) support vector machines with linear kernel (svmLinear), (b) support vector machines with radial basis kernel (svmRadial), (c) k-nearest neighbors (KNN), (d) neural network (NNET), and (e) deep neural network (DNN).

The same sampling of 5-fold CV as RF is used for all other compared algorithms. And the same measure of the pathway-based classification performance, *G* (described in Section III-A), is used to compare all algorithms. The five typical classifiers are implemented through R package *caret* version 6.0-81 [36]. Parameters of classifiers are tuned through 5 different unique parameter combinations in the train function of caret. DNN is implemented through R package *h2o* version 3.20.0.8 [37], where the default parameters are used.

TABLE I
REAL DATASETS EMPLOYED IN THE PAPER

Dataset	No. Cells	No. Cell Populations	No. Genes	Organism
Grun	160	2	12535	Mus musculus
Kolod	704	3	38576	Mus musculus
Patel	430	5	5948	Homo sapiens
Sharma	1302	13	22744	Homo sapiens

We compare our pathway-based RF clustering to other state-of-the-art clustering methods including *RAFSIL1/2* [14], [15], which also use RF for similarity learning through newly constructed feature space from PCA and k-means clustering of genes; *SIMLR* [11], which assesses cellular similarity based on weights from multiple kernels, and *SC3* [12], which is a consensus method integrating correlation distance, PCA, k-means and cluster-based similarity partitioning. The widely used measure calculated from biological clusters and predicted clusters, adjusted rand index (ARI) [38], is used to compare the clustering performance. Default parameters of those methods are used.

E. Simulation Study

To better understand the pathway-based classification performance of RF in differentiating cell populations for scRNA-Seq data, we have done a simulation study based on data from real scRNA-Seq data. Two cases of hypotheses are simulated: the null hypothesis where a pathway has no informative gene in differentiating cell populations, and the alternative hypothesis where a pathway has certain informative genes in relation to different cell populations.

For the null case, we randomly select genes with non-positive MDA (i.e., non-informative genes for the classification) from pathways with the lowest ranked *G* to create simulated pathways of different sizes (number of genes involved in): 10, 20, 40, and 80. For each simulated pathway size, different number of cells in three cell populations: 50 vs 50 vs 50, 100 vs 100 vs 100, and 150 vs 150 vs 150 are randomly selected and permutated. With each combination of pathway size and number of cells, we simulate 100 times, i.e., 100 different simulated pathways with each combination. *G* in (4) is recorded in each iteration of the simulation.

For the alternative case, besides pathway size and number of cells, we add a variation of % informative genes in a pathway (20%, 40%, and 60%) to evaluate their impact on the performance measured by *G*. Informative genes are randomly selected from genes with positive MDAs in top 10 ranked pathways from both BioCarta and KEGG databases.

IV. RESULTS

A. Real Datasets

To compare the performance of pathway-based RF classification to other classifiers for scRNA-Seq data, four real publicly available datasets are utilized. Table I summarizes the basic information of the four datasets. Grun *et al.* [39] highlighted the variation of transcriptomes of mouse embryonic stem cells (mESCs) cultured in two different conditions in their study. Kolodziejczyk *et al.* [40] demonstrated distinct levels of gene expression of mESCs grown under three different conditions. Patel *et al.* [7] demonstrated intratumoral heterogeneity in five different glioblastoma tumors. Sharma *et al.* [41] investigated the phenotypic heterogeneity and homogeneity of thirteen lines of tumor cells with different drug-resistant/holiday models. Additional information of the four datasets including normalization methods, sparsity of the matrices, and download sites is listed in Supplementary Table I.

B. Application of Proposed Approach

In this section, we present the results of applications of our proposed approach. The first subsection presents the application of pathway-based RF classification to find discriminative pathways and to construct GGIs network for Kolod dataset using BioCarta pathways. The application results for Patel dataset using KEGG pathways are shown in Supplementary Results R1. The second subsection presents the application of pathway-based RF clustering for cell populations.

Two parameters can be tuned in RF, which are the number of features to be randomly selected in the node splitting (*mtry*), and the total number of trees in the forest (*ntree*). In our approach, we apply the default setting of $mtry$ = square root of input number of features, which is suggested to give optimal results [19]. And we set $ntree = 500$ as the default value since further incrementing *ntree* has little improvement on the classification accuracy. On the other hand, the computational time increases proportionally with the increment of *ntree*. Fig. 1(a) shows the effect of *ntree* to the classification accuracy among all available pathways. Fig. 1(b) shows the relationship between *ntree* and computational time in minutes.

1) Identification of Discriminative Pathways and the Construction of GGIs Networks: We apply our proposed approach to identify BioCarta pathways that are good at explaining the heterogeneity of mESCs grown in three different conditions: serum (lif), two inhibitors (2i), and the alternative ground state (a2i) in the Kolod dataset. 217 BioCarta pathways that have pathway size no less than 5 are investigated. Pathways are prioritized based on the goodness score G in (4), which is the average pathway-based classification accuracy of 5-fold CV.

Table II shows the top 10 discriminative BioCarta pathways in the Kolod dataset. The original study of the Kolod dataset found that most heterogeneous genes of cells in different conditions were related to cell cycle, MAPK signaling, and basic metabolism [40]. Consistently, the top 10 discriminative BioCarta pathways identified by our approach include the pathway related to MAPK signaling, pathways related to cell cycle such as TCR Pathway, G1 Pathway, RACCYCD Pathway, and P53 Pathway, as well as pathways related to basic metabolism such as PPARA Pathway and Free Pathway. In the original study of Kolod, cells within a2i population were inhibited with glycogen synthase kinase 3 (GSK3) [40]. And our approach has successfully identified GSK3 Pathway as a good classifier.

TABLE II
TOP 10 DISCRIMINATIVE BIOCARTA PATHWAYS IN KOLOD

Pathway	G [a]	Pathway Size
TCR Pathway	0.8721	44
PPARA Pathway	0.8594	55
MAPK Pathway	0.8579	86
G1 Pathway	0.8452	27
EDG1 Pathway	0.8366	27
Integrin Pathway	0.8310	38
Free Pathway	0.8295	9
RACCYCD Pathway	0.8211	26
GSK3 Pathway	0.8209	27
P53 Pathway	0.8196	15

[a] G is the average pathway-based classification accuracy of 5-fold CV that is calculated based on (3) and (4).

The RF based GGIs network for important genes in discriminative BioCarta pathways in the Kolod dataset is shown in Fig. 2. Each node represents an important gene selected based on MDA. Size and the color of the node are related to the total number of discriminative pathways a gene is involved in. ‘Cross-talk’ genes linking multiple discriminative pathways are highlighted and their influences on classification are shown in Supplementary Results R2. For the Kolod dataset, Mapk3, also known as Erk1 links 6 discriminative pathways: TCR, PPARA, MAPK, EDG1, Integrin, and RACCYCD, which are pathways mostly related to cell cycle. The layout of the network depends on the co-occurrence matrix C of (9) indicating the co-occurrence of genes in discriminative pathways. Co-functional genes sharing a greater number of pathways are clustered. For instance, genes Cdk2, Cdkn1a, and Cdk6 belong to pathways G1 and RACCYCD are clustered in the network. The weight of the undirected edge between two genes is the prevalence of GGIs in RF as defined in (6), (7) and (8). In total 29 nodes are in the network and the default calibration of graph density D = 0.05, 49 edges between genes with top ranked prevalence of interaction in RF are created. Fig. 3 shows patterns of expression level in different cell populations of gene pairs having top 2 prevalent interaction in RF. In general, these interacting genes in RF together give a relatively good separation of cell populations and may correlate with each other. For example, genes Gpx1 and Gja1 have a significant Spearman’s correlation (p value < 0.0001, coefficient $= -0.60$) and both are heterogeneously expressed in three cell populations. To assess the feasibility and stability of the I values, we illustrate the relationship between the number of trees in RF and the I values in the Supplementary Fig. 2 for the top 10 and the bottom 10 gene-gene pairs ranked based on the default RF. The I values converges to a stable value with growing number of trees.

2) Application of Pathway-Based RF Clustering: We apply our pathway-based RF clustering for predicting cell populations for the Kolod and Patel datasets using BioCarta pathways and KEGG pathways respectively. Fig. 4(a) and Fig. 4(b) are 2-dimensional (2D) visualization of the predicted cell populations and biological cell populations for Kolod and Patel datasets by our method. Here we select the 10 most relevant pathways for clustering. Fig. 4(c) and Fig. 4(d) present the Venn diagrams of

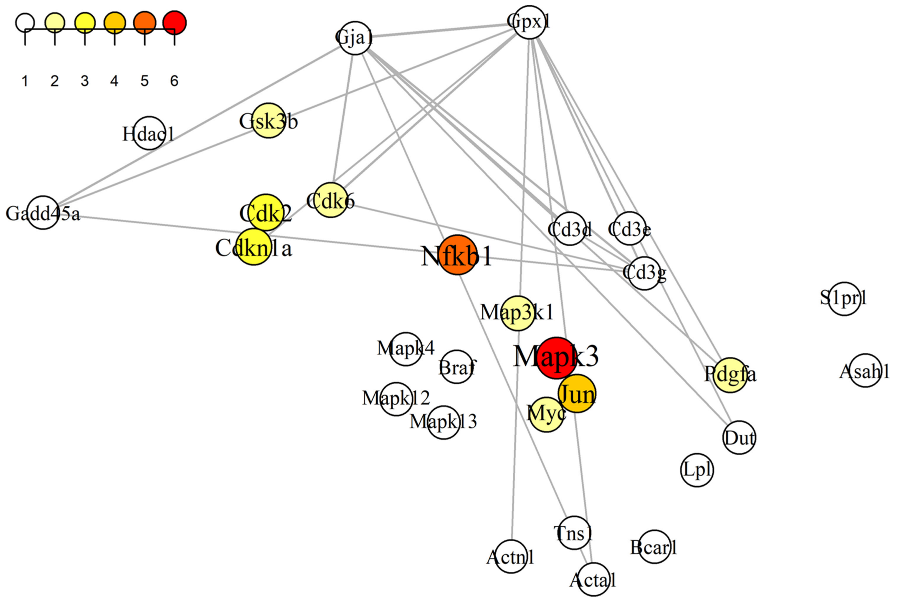

Fig. 2. The GGIs network for important genes from discriminative BioCarta pathways in the Kolod dataset. Each node represents a gene. Layout of nodes depends on co-occurrences of genes in discriminative pathways. Size and color of a node indicate number of discriminative pathways the gene involved in. An edge is created if the prevalence of interaction between two genes in RF is high given the calibration of graph density D = 0.05.

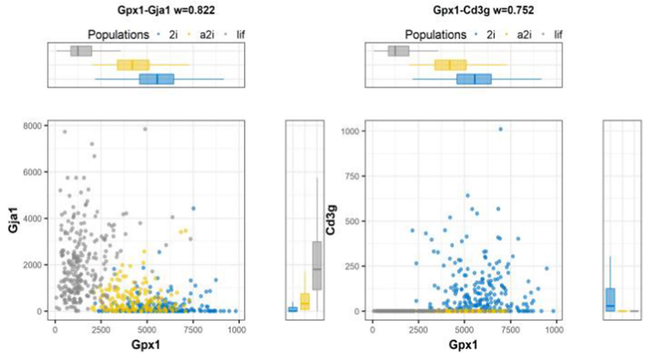

Fig. 3. Expression values of top 2 prevalent interacting gene pairs in the Kolod dataset. Genes of cells are grouped by cell populations and shown in colors, and ‘w’ in the title represents the prevalence of interaction between two genes.

10 relevant pathways selected from clustering and 10 discriminative pathways selected from classification.

C. Comparison of Classification Methods and Clustering Methods

1) Comparison of Pathway-Based RF Classification to Other Methods: We compare the pathway-based classification performance in differentiating cellular heterogeneity of different classifiers. Table III summarizes the average classification accuracy of 5-fold CV of different classifiers based on both BioCarta pathways and KEGG pathways for four real scRNA-Seq datasets. A non-informative classifier (NIC) that assigns classes of all observations as the majority class is also included in Table III for the comparison. The distribution of classification accuracy of different classifiers for four datasets through density curves

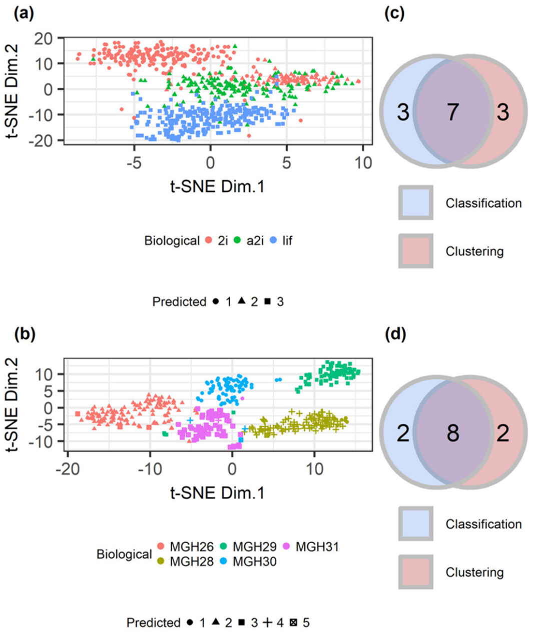

Fig. 4. Visualization of cell populations (biological vs predicted) through tSNE from relevant pathways and the overlap of selected pathways between clustering and classification. (a) 2D Visualization of cell populations in Kolod. (b) 2D Visualization of cell populations in Patel. (c) Venn diagram of selected BioCarta pathways between clustering and classification in Kolod. (d) Venn diagram of selected KEGG pathways between clustering and classification in Patel.

based on BioCarta and KEGG pathways are shown in Fig. 5(a) and Fig. 5(b), respectively. Non-informative rate (NIR) in titles of subfigures represents the classification accuracy of NIC, which equals to the ratio of the majority population of cells over all cells.

TABLE III
COMPARISON OF AVERAGE CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFIERS BASED ON BIOCARTA AND KEGG PATHWAYS

Classifiers	Grun	Kolod	Patel	Sharma
RF	0.7954	0.7326	0.5540	0.5421
KNN	0.7453	0.6172	0.4676	0.3709
NNET	0.7814	0.5684	0.4723	0.2480
svmLinear	0.7806	0.6720	0.5079	0.4647
svmRadial	0.7718	0.6530	0.5422	0.4070
DNN	0.7895	0.6083	0.4200	0.3611
NIC*	0.5000	0.4190	0.2744	0.1943

*NIC is the represents a non-informative classifier that assigns classes of all observations as the majority class.

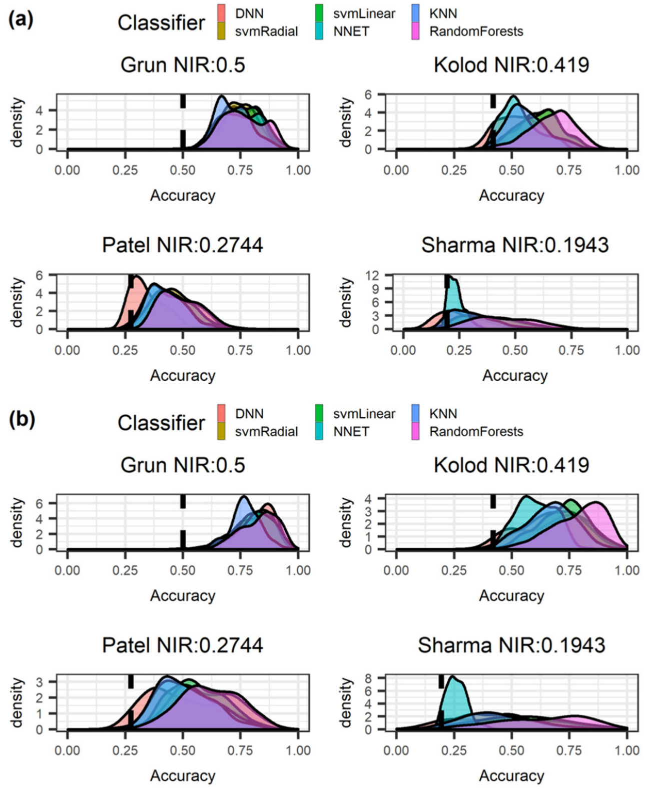

Fig. 5. The distribution of classification accuracy of different classifiers for 4 datasets. (a) Density curves of classification accuracy based on BioCarta pathways. (b) Density curves of classification accuracy based on KEGG pathways. NIR: non-informative rate. The vertical line in each plot indicates NIR to the dataset.

TABLE IV
COMPUTATIONAL TIME (MINUTES) OF DIFFERENT CLASSIFIERS

Classifiers	Grun	Kolod	Patel	Sharma
RF	1.87	17.85	5.71	53.45
KNN	8.79	11.50	8.19	14.65
NNET	22.90	77.65	34.00	200.50
svmLinear	9.15	76.85	8.40	184.50
svmRadial	11.15	21.90	15.20	134.50
DNN	29.00	73.90	39.75	127.00

TABLE V
COMPARISON OF PERFORMANCE (ARI) OF CLUSTERING

Clustering Method	Grun	Kolod	Patel	Sharma
scPathwayRF	0.8781	0.9880	0.8928	0.7945
RAFSIL1	0.8547	1.0000	0.9682	0.6535
RAFSIL2	0.5599	0.9960	0.9698	0.6974
SIMLR	0.4696	1.0000	0.8089	0.5719
SC3	0.8547	1.0000	0.9891	0.7666

Results from Table III and Fig. 5 indicate that RF outperforms other classifiers in pathway-based classification of different cell populations for scRNA-Seq data. Results also show the average classification accuracies of our classifiers are higher than the baseline NIR which represents the accuracy of random guessing accounting for class sizes. Given the baseline NIR, the average classification accuracies of datasets with larger number of classes are generally lower than those with fewer classes. Table IV shows the average computational time of different classifiers to investigate all available pathways in databases: 186 BioCarta pathways, and 217 KEGG pathways. Only KNN requires less computational time than RF in datasets containing more than three classes (5 classes in Patel and 13 classes in Sharma). In terms of computational time, pathway-based RF for scRNA-Seq data is especially efficient for classifying multiple cell populations. Support vector machine with either linear kernel or radial kernel slows down notably when the number of cells and populations of cells both increase. The results of a sensitivity analysis to assess the influence of tuning parameters on the performance of different classifiers is provided in Supplementary Results R3, which also shows decent performance of RF with default parameters.

2) Comparison of Pathways-Based RF Clustering to Other Methods: The performance of our clustering approach is compared to other state-of-the-art clustering methods by ARI metric. Table V summarizes the performance of different methods for tested datasets. In general, *SC3* outperforms others and *SIMLR* appears to be the least competitive. Although our proposed method has the lowest performance for the Kolod dataset and the second lowest performance for the Patel dataset, it outperforms others for the Grun and Sharam datasets. Regarding to the computational time, our method is comparable with *RAFSIL1/2* and costs around 50% more computing time compared to *SIMLR* and *SC3*.

D. Results of Simulation

As described in Section III-D, datasets with different combinations of pathway size, number of cells (sample size), and % informative genes are simulated to understand the pathway-based classification performance of RF for scRNA-Seq data. The mean and standard deviation of the pathway-based classification performance, with 100 simulations for each combination are summarized in Table VI and Table VII.

Table VI shows classification performance with different combinations of pathway size and sample size under the null hypothesis in which the investigated pathway is non-informative

TABLE VI
SIMULATIONS UNDER THE NULL MEASURED BY G^* (MEAN ± SD)

Pathway Size	Number of Cells		
	50 vs 50 vs 50	100 vs 100 vs 100	150 vs 150 vs 150
10	0.33 ± 0.05	0.33 ± 0.03	0.33 ± 0.03
20	0.33 ± 0.04	0.33 ± 0.04	0.33 ± 0.03
40	0.34 ± 0.04	0.33 ± 0.04	0.33 ± 0.03
80	0.32 ± 0.05	0.33 ± 0.03	0.33 ± 0.03

$^*G$ is the average pathway-based classification accuracy of 5-fold CV that calculated based on (3), (4). SD, standard deviation.

TABLE VII
SIMULATIONS UNDER THE ALTERNATIVE MEASURED BY G^* (MEAN ± SD)

Number of Cells 50 vs 50 vs 50			
Pathway Size	% Informative Genes in the Pathway		
	20	40	60
10	0.58 ± 0.06	0.70 ± 0.06	0.76 ± 0.06
20	0.69 ± 0.07	0.81 ± 0.05	0.86 ± 0.04
40	0.80 ± 0.05	0.90 ± 0.03	0.94 ± 0.02
80	0.89 ± 0.03	0.95 ± 0.02	0.97 ± 0.01
Number of Cells 100 vs 100 vs 100			
Pathway Size	% Informative Genes in the Pathway		
	20	40	60
10	0.59 ± 0.07	0.70 ± 0.06	0.77 ± 0.05
20	0.70 ± 0.06	0.83 ± 0.04	0.87 ± 0.03
40	0.83 ± 0.04	0.91 ± 0.02	0.94 ± 0.02
80	0.91 ± 0.03	0.96 ± 0.01	0.98 ± 0.01
Number of Cells 150 vs 150 vs 150			
Pathway Size	% Informative Genes in the Pathway		
	20	40	60
10	0.61 ± 0.07	0.70 ± 0.06	0.78 ± 0.05
20	0.71 ± 0.05	0.83 ± 0.04	0.89 ± 0.03
40	0.83 ± 0.04	0.92 ± 0.02	0.95 ± 0.01
80	0.92 ± 0.02	0.97 ± 0.01	0.98 ± 0.01

$^*G$ is the average pathway-based classification accuracy of 5-fold CV that calculated based on (3), (4). SD, standard deviation.

to differentiate cell populations. The classification performance under the null hypothesis is close to a random guessing (0.33) for the simulated dataset containing 3 balanced classes.

Table VII shows classification performance under the alternative hypothesis with different combination of % informative genes in the investigated pathway. Under the alternative hypothesis where the investigated pathway is somehow informative in differentiating cell populations, the classification performance is better than a random guess. And as the % informative genes in a pathway, pathway size, or sample size increases, the classification performance will be improved as well.

V. DISCUSSION AND CONCLUSION

As scRNA-Seq has emerged in recent years, corresponding computational analytic approaches and user-friendly packages are needed. Although numerous machine learning based unsupervised methods have been proposed for scRNA-Seq data, assigning heterogeneous biological functions of determined cell populations is still a challenging issue. Appreciating the importance of this fact, in this study, we present a pathway-based classification approach to find the heterogeneous functional pathways that differentiate cell populations using RF. The robustness of RF can be extended to scRNA-Seq data that are noisier and sparser compared to bulk RNA-Seq data. Advantages of RF compared to other classifiers in terms of both classification performance and computational time are demonstrated. With its ensemble nature, RF is robust to relatively noisy data and has great performance for high-dimensional data [42]. In addition, RF provides decent performance with default parameters without the need for tuning in classifying cell populations in the pathway-based setting. Through simulation study, we have confirmed our pathway-based RF classification approach for scRNA-Seq data is able to distinguish between the true informative pathways (the alternative case) and the non-informative ones (the null case).

We also provide a clustering method for scRNA-Seq that combines both gene-based and pathway-based information using RF. The performance and computational time of our method is comparable to *RAFSIL1/2* [14], [15], which also uses RF for similarity learning for scRNA-Seq. Both methods perform better than *SIMLR* [11]. Although *SC3* [12] generally outperforms other methods, our method is at the top for two datasets and identifies relevant pathways for clustering that can help users to infer biological functions of cell clusters. Additionally, clustering result of other methods can be applied into our pipeline for prioritizing discriminative pathways and constructing GGIs network to infer functional heterogeneity between cell clusters.

We further propose a novel approach to construct GGIs networks that can identify gene pairs that interact together and can also serve as strong predictors of cell populations. The co-occurrence of genes in different discriminative pathways and 'cross-talk' genes connecting them are also highlighted in our networks. These networks can facilitate researchers to pinpoint important genes contributing to pathway cross-talks and to easily capture co-functional clusters among multiple discriminative pathways.

Our work addresses the importance of functional pathways in understanding heterogeneous biological function in different cell populations. Our novel networks not only group co-functional genes with common discriminative pathways via co-occurrence matrix but also allow the identification of GGIs in differentiating cell populations. Despite the advantages mentioned above, the proposed pathway-based analytic framework also has some limitations. Building multiple pathway-based RF models requires more computational resources compared to testing individual genes followed by functional enrichment analysis as well as compared to clustering cells based on purely gene expression profiles. However, with today's computing power of general PCs the computational time of the analysis may not be an issue; if it is, pathway-based analysis can easily be distributed across computation cores. Another limitation is that constructing pathway-based models depends on whether the genes are annotated, genes that are not mapped to pathways are overlooked. However, pathways do provide biological context and can enhance our understanding of heterogeneous cellular functions using scRNA-Seq. Moreover, the pathway databases are also updated frequently. In conclusion, our work can stimulate more future research that incorporates prior biological knowledge in the analysis of scRNA-Seq data in the era of precision health and medicine.

References

[1] F. Tang *et al.*, “mRNA-Seq whole-transcriptome analysis of a single cell,” *Nat. Methods*, vol. 6, no. 5, pp. 377–382, May 2009.

[2] B. Treutlein *et al.*, “Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq,” *Nature*, vol. 509, no. 7500, pp. 371–375, May 2014.

[3] L. Yan *et al.*, “Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells,” *Nat. Struct. Mol. Bio.*, vol. 20, no. 9, pp. 1131–1139, Sep. 2013.

[4] F. Buettner *et al.*, “Computational analysis of cell-to-cell heterogeneity in single-cell RNA sequencing data reveals hidden subpopulations of cells,” *Nat. Biotechnol.*, vol. 33, no. 2, pp. 155–160, Feb. 2015.

[5] A. C. Villani *et al.*, “Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors,” *Science*, vol. 356, no. 6335, pp. eaah4573, Apr. 2017.

[6] P. Dalerba *et al.*, “Single-cell dissection of transcriptional heterogeneity in human colon tumors,” *Nat. Biotechnol.*, vol. 29, no. 12, pp. 1120–1127, Nov. 2011.

[7] A. P. Patel *et al.*, “Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma,” *Science*, vol. 344, no. 6190, pp. 1396–1401, Jun. 2014.

[8] J. Žurauskienė and C. Yau, “pcaReduce: Hierarchical clustering of single cell transcriptional profiles,” *BMC Bioinf.*, vol. 17, p. 140, Mar. 2016.

[9] E. L-A. D. Amir *et al.*, “viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia,” *Nat. Biotechnol.*, vol. 31, no. 6, pp. 545–552, Jun. 2013.

[10] L. Haghverdi, F. Buettner, and F. J. Theis, “Diffusion maps for high-dimensional single-cell analysis of differentiation data,” *Bioinf.*, vol. 31, pp. 2989–2998, May 2015.

[11] B. Wang *et al.*, “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,” *Nat. Methods*, vol. 14, no. 4, pp. 414–416, Apr. 2017.

[12] V. Y. Kiselev *et al.*, “SC3: Consensus clustering of single-cell RNA-seq data,” *Nat. Methods*, vol. 14, pp. 483–486, May 2017.

[13] X. Li and K. Wong, “Single-cell RNA-seq data interpretation by evolutionary multiobjective clustering,” *IEEE Trans. Comput. Biol. Bioinform.*, Mar. 2019, Epub, doi: 10.1109/TCBB.2019.2906601.

[14] M. B. Pouyan and M. Nourani, “Clustering single-cell expression data using random forest graphs,” *IEEE J Biomed Health Inform.*, vol. 21, no. 4, pp. 1172–1181, Jul. 2017.

[15] M. B. Pouyan and D. Kostka, “Random forest based similarity learning for single cell RNA sequencing data,” *Bioinformatics*, vol. 34, no. 13, pp. i79–i88, Jul. 2018.

[16] D. A. Lawson *et al.*, “Tumour heterogeneity and metastasis at single-cell resolution.,” *Nat. Cell Biol.*, vol. 20, no. 12, pp. 1349–1360, Nov. 2018.

[17] L. Zhang and S. Zhang, “Comparison of computational methods for imputing single-cell RNA-sequencing data,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, Epub, Jun. 2018, doi: 10.1109/TCBB.2018.2848633.

[18] J.K. Kim *et al.*, “Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression,” *Nat. Commun.*, vol. 6, p. 8687, Oct. 2015.

[19] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[20] R Díaz-Uriarte and S. A. de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, p. 3, Jan. 2006.

[21] J.W. Lee, J. B. Lee, M. Park, and S. K. Song, “An extensive comparison of recent classification tools applied to microarray data,” *Comput. Stat. and Data Anal.*, vol. 48, no. 4, pp. 869–885, Apr. 2005.

[22] A. Statnikov, L. Wang, and S. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” *BMC Bioinf.*, vol. 9, p. 319, Jul. 2008.

[23] H. Pang *et al.*, “Pathway analysis using random forests classification and regression,” *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, Aug. 2006.

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth International Group, 1984.

[25] M. Kanehisa *et al.*, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.

[26] L. Breiman, *Manual on Setting up, Using, and Understanding Random Forests v4. 0.*, CA, USA: Statistics Department University of California Berkeley, 2003.

[27] T. Shi and S. Horvath, “Unsupervised learning with random forest predictors,” *J. Comput. Graphical Statistics*, vol. 15, no. 1, pp. 118–138, Mar. 2006.

[28] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis,” *J. R. Stat. Soc. B.*, vol. 61, no. 3, pp. 611–622, Sep. 1999.

[29] L. J. P. van der Maaten and G. E. Hinton, “Visualizing high-dimensional data using t-SNE,” *JLMR.*, vol. 9, pp. 2579–2605, 2008.

[30] J. H. Krijthe, “Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation,” R package version 0.15., Nov. 2018. [Online]. Available: https://cran.r-project.org/web/packages/Rtsne/index.html

[31] C. Hennig, “fpc: Flexible Procedures for Clustering,” R package version 2.2-2, Jun. 2018. [Online]. Available: https://cran.r-project.org/web/packages/fpc/index.html

[32] R. L. Thorndike, “Who belongs in the family?,” *Psychometrika.*, vol. 18, no. 4, pp. 267–276, Dec. 1953.

[33] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, pp. 18–22, Dec. 2002. R Package Version 4.6-14, March 25, 2018. [Online]. Available: https://cran.r-project.org/package=randomForest

[34] S. Basu, K Kumbier, J.B. Brown, and B. Yu, “Iterative random forests to discover predictive and stable high-order interactions,” *PNAS*, vol. 115, no. 8, pp. 1943–1948. Feb. 2018. R Package Version 2.0.0, July 27, 2017. [Online]. Available: https://cran.r-project.org/package=iRF

[35] G. Csárdi G. and T. Nepusz, “The igraph software package for complex network research,” *Inter J.*, vol. Complex Systems 1695, pp. 1–9, 2006. R Package Version 1.2.2, July 27, 2018. [Online]. Available: https://cran.r-project.org/src/contrib/Archive/igraph/igraph_1.2.2.tar.gz

[36] M. Kuhn, “Building predictive models in R using the caret package,” *J. Stat. Softw.*, vol. 28, pp. 1–26, Nov. 2008. R Package Version 6.0-81, Nov 20, 2018. [Online]. Available: https://cran.r-project.org/package=caret

[37] The H2O.ai Team, “h2o: R Interface for H2O,” R Package Version 3.20.0.8, Sep. 25 2018. [Online]. Available: https://cran.r-project.org/src/contrib/Archive/h2o/h2o_3.20.0.8.tar.gz

[38] L. Hubert and P. Arabie, “Comparing partitions,” *J. Classification*, vol. 2, no. pp. 193–218, 1985.

[39] D. Grün, L. Kester, and A. van Oudenaarden, “Validation of noise models for single-cell transcriptomics,” *Nat. Methods*, vol. 11, no. 6, pp. 637–640, Jun. 2014.

[40] A. A. Kolodziejczyk *et al.*, “Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation,” *Cell Stem Cell.*, vol. 17, no. 4, pp. 471–485, Oct. 2015.

[41] A. Sharma *et al.*, “Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy,” *Nat Commun.*, vol. 9, p. 4931, Nov. 2018.

[42] R. Caruana, N. Karampatziakis, and A. Yessenalina, “An empirical evaluation of supervised learning in high dimensions,” in *Proc. 25th Int’l Conf. Machine Learn. (ICML ’08)*, 2008.