Published online 8 August 2018

Nucleic Acids Research, 2018, Vol. 46, No. 21 ***e124***
doi: 10.1093/nar/gky698

CellBIC: bimodality-based top-down clustering of single-cell RNA sequencing data reveals hierarchical structure of the cell type

Junil Kim [1,2,3], Diana E. Stanescu [1,4,*] and Kyoung Jae Won[1,2,3,*]

[1]Institute for Diabetes, Obesity and Metabolism, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [2]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA, [3]Biotech Research and Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen, Denmark and [4]Division of Endocrinology and Diabetes, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

Received March 07, 2018; Revised July 19, 2018; Editorial Decision July 22, 2018; Accepted July 23, 2018

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) is a powerful tool to study heterogeneity and dynamic changes in cell populations. Clustering scRNA-seq is essential in identifying new cell types and studying their characteristics. We develop CellBIC (single Cell BImodal Clustering) to cluster scRNA-seq data based on modality in the gene expression distribution. Compared with classical bottom-up approaches that rely on a distance metric, CellBIC performs hierarchical clustering in a top-down manner. CellBIC outperformed the bottom-up hierarchical clustering approach and other recently developed clustering algorithms while maintaining the hierarchical structure of cells. Importantly, CellBIC identifies type 2 diabetes and age specific β cell signatures characterized by *SIX3* and *CDH2*, respectively.

INTRODUCTION

The advent of high capacity single cell RNA-seq (scRNA-seq) allows the characterization of large number of single cell transcriptomes (1–6). Clustering is an important step to study cell heterogeneity and characterize previously unknown cell sub-population. For clustering cell-to-cell distance is usually used to identify cells with similar expression profiles. A classical hierarchical clustering method uses the cell-to-cell distance to reconstruct a tree structure, where each branch represents a sub-type (2–10). Accordingly, the choice of an appropriate metric and a linkage criterion influences the clustering results. Also, a hierarchical clustering method often fails to find a large group in the process of collecting small groups to reconstruct a tree structure due to the conflict between similarity across multiple groups (11,12). Ensemble approaches have been developed for scRNA-seq clustering to compensate biases from using a metric or a criterion (13,14). SC3 (13) uses a consensus of k-means clustering results followed by hierarchical clustering. SIMLR (14) provides a similarity metric learned by combining multiple kernels. Even though successful, these approaches still rely on the choice of a metric and a linkage criterion.

Dimension reduction approaches such as principle component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) have been widely applied for clustering in order to handle multi-dimensional scRNA-seq data efficiently by visualizing cells in a two or three dimensional space (2,4,5,6,8–10,15–20). Subsequently, clustering is performed by grouping cells based on the proximity in the reduced dimension. However, these approaches can fail to identify cell types when cells are intermingled in the reduced domain.

Here, we develop CellBIC (single Cell BImodal Clustering), a novel clustering approach for scRNA-seq. Compared with a classical 'bottom-up' hierarchical clustering approach that builds up a tree from a distance matrix, CellBIC implements a 'top-down' approach by performing clustering by dividing the datasets recursively to reconstruct a hierarchical structure (21). While a top-down hierarchical clustering method requires additional computation time for identifying an optimal division point (22), it can provide a better clustering performance and interpretation than bottom-up algorithms (21–25). A top-down clustering equips with a simple flat clustering to divide datasets into large sub-groups (22). Often, a K-means algorithm has been used as a flat clustering due to its execution speed (22). Instead of using a K-means algorithm, we designed a new flat clustering approach for scRNA-seq data using multi-modal patterns in gene expression.

Multi-modality is an intrinsic characteristic observed in heterogeneous scRNA-seq data. For instance, if the expression levels of a gene are high in a cell type and low in other

*To whom correspondence should be addressed. Tel: +45 3533 1419; Email: kyoung.won@bric.ku.dk
Correspondence may also be addressed to Diana E. Stanescu. Tel: +1 267 648 2541; Fax: +1 215 590 8033; Email: stanescud@email.chop.edu

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

cell type, its distribution in the mixed population will show a bimodal distribution in scRNA-seq. Bimodality in expression is also previously observed in the scRNA-seq from mouse bone-marrow-derived dendritic cells due to the potential cell specificity of a sub-population in responding to lipopolysaccharide (26).

CellBIC identifies cellular clusters by considering a bimodal expression pattern in a top-down manner. By dissecting cells based on bimodal memberships recursively, CellBIC reconstructs a hierarchical tree structure. We applied CellBIC to various scRNA-seq data including human pancreas (3,6), mouse cortex (4), and mouse lung (5). CellBIC shows a superior performance for clustering of scRNA-seq data over traditional bottom-up clustering algorithms as well as other state-of-the-art clustering algorithms. The hierarchical structure that CellBIC identified reconstructed the developmental lineages of the pancreas. Applying CellBIC further to pancreatic β cells, we identified genes in β cells associated with type 2 diabetes (T2D) and aging.

MATERIALS AND METHODS

Clustering evaluation

t-SNE (27) and density-based spatial clustering of applications with noise (DBSCAN) (28) were implemented in MATLAB 2017b. Bottom-up hierarchical clustering was performed by two MATLAB functions 'linkage' and 'dendrogram'. For t-SNE and bottom-up clustering, we used genes whose log2 transformed TPM >1 in >25% cells and the coefficient of variation of the $\log_2$ transformed TPM >1. For SC3 and SIMLR, we used their default parameters.

To benchmark clustering algorithms, each obtained cluster was assigned to a cell type with the best matching cell type. By comparing the assigned cell type with the gold standard, we calculated adjusted rand Index (ARI) for each clustering result. We considered the maximum ARI of the clustering results for different numbers of clusters and for five repeats as the ARI of each clustering algorithm. ARI was calculated by a MABLAB function 'rand_index' in a package Adjusted Rand Index version 1.0 with 'adjusted' option using MATLAB 2017b (https://www.mathworks.com/matlabcentral/fileexchange/49908-adjusted-rand-index).

CellBIC uses bimodal distribution in scRNA-seq for clustering

A multi-modal distribution in expression is often observed in scRNA-seq data from a mixed cell population. For instance, insulin (*INS*) expression levels in the scRNA-seq from human pancreas showed a bimodal distribution: the higher mode for β cells (*INS*+) the lower mode for other cells (*INS*-) regardless of the donors (Supplementary Figure S1). This suggests a potential use of modality for scRNA-seq clustering.

Hypothesizing that cells that belong to a mode in a bimodal distribution consistently across multiple genes share similarity, we designed a top-down hierarchical clustering approach. To implement clustering, we obtain Boolean membership of cells (High/Low or 0/1) after fitting the distribution of expression levels of each gene with a Gaussian mixture model (Figure 1A). To obtain bimodal distribution, we used a Gaussian mixture model with two modes. We only consider a gene as a candidate based on t-statistics ($t > 10$). In CellBIC, genes that do not make cells belong to a membership are discarded (Figure 1B). For this, CellBIC calculates the Hamming distance using the Boolean membership. The Hamming distance (H_{ij}) of two Boolean vectors i and j is defined by the ratio of the number of mismatch. The absolute Hamming distance (AH_{ij}) between i and j is defined as follows:

$$AH_{ij} = \begin{cases} H_{ij} & \text{if } H_{ij} < 0.5 \\ 1 - H_{ij} & \text{otherwise} \end{cases}$$

In this configuration, the Hamming distance of 1 is for the perfect mismatch for a gene pair and 0.5 for a random match. Two genes with maximum Hamming distance are selected as the top seed gene pair. We limited that the seed gene pair has a distance larger than 0.7 to remove seed gene pairs weakly mismatching each other. We further confirmed that the performance was robust to the Hamming distance cutoff for a seed gene pair (Supplementary Figure S2). Then, we selected genes showing consistent or contrasting Boolean membership with the selected two seed genes. A gene is added if average absolute Hamming distance with the existing genes is below 0.2 (correlation test P-value <0.01, and binomial test P-value < 1e–25), which provided robust clustering performances for several independent scRNA-seq datasets (Supplementary Figure S3).

The collective membership will show two groups of cells aligned with the seed gene pair (Figure 1C).

Dissecting cells using the hamming code table

The Hamming codes for the aligned genes were represented in the membership matrix. The matrix is composed of two gene groups aligned to the two seed genes, respectively. The Boolean values were switched (0 to 1 and 1 to 0) when appropriate. The maximum of the moving standard deviation (window = 1/15 of the total number of cells) on the column sum of the matrix determines the dissection point of the two groups of cells. Consequently, the dissecting point and the genes aligned to the two seed genes form four quadrants in the membership matrix. The majority values of the two crossing quadrants are same. For instance, the majority value of the second and the fourth quadrants in Figure 1C is 0 and the other two quadrants (first and third quadrants in Figure 1C) is 1. If the majority component is not observed in an alternative manner from the first to the fourth quadrant, dissection is discarded and the clustering stops. To prevent CellBIC from falling into a local minimum, this procedure is repeated using the top 5 seed genes (Supplementary Figure S4). When selecting the aligned gene sets with the seed genes, genes used for previous dissection are not considered for the alignment with the next seed genes. Genes that provide best dissection among them is selected based on the following clustering score C.

$$C = \left(\frac{E}{4} - 0.7\right) N$$

where E denotes the sum of the ratios of '1' in the second and the fourth quadrants and the ratios of '0' in the first

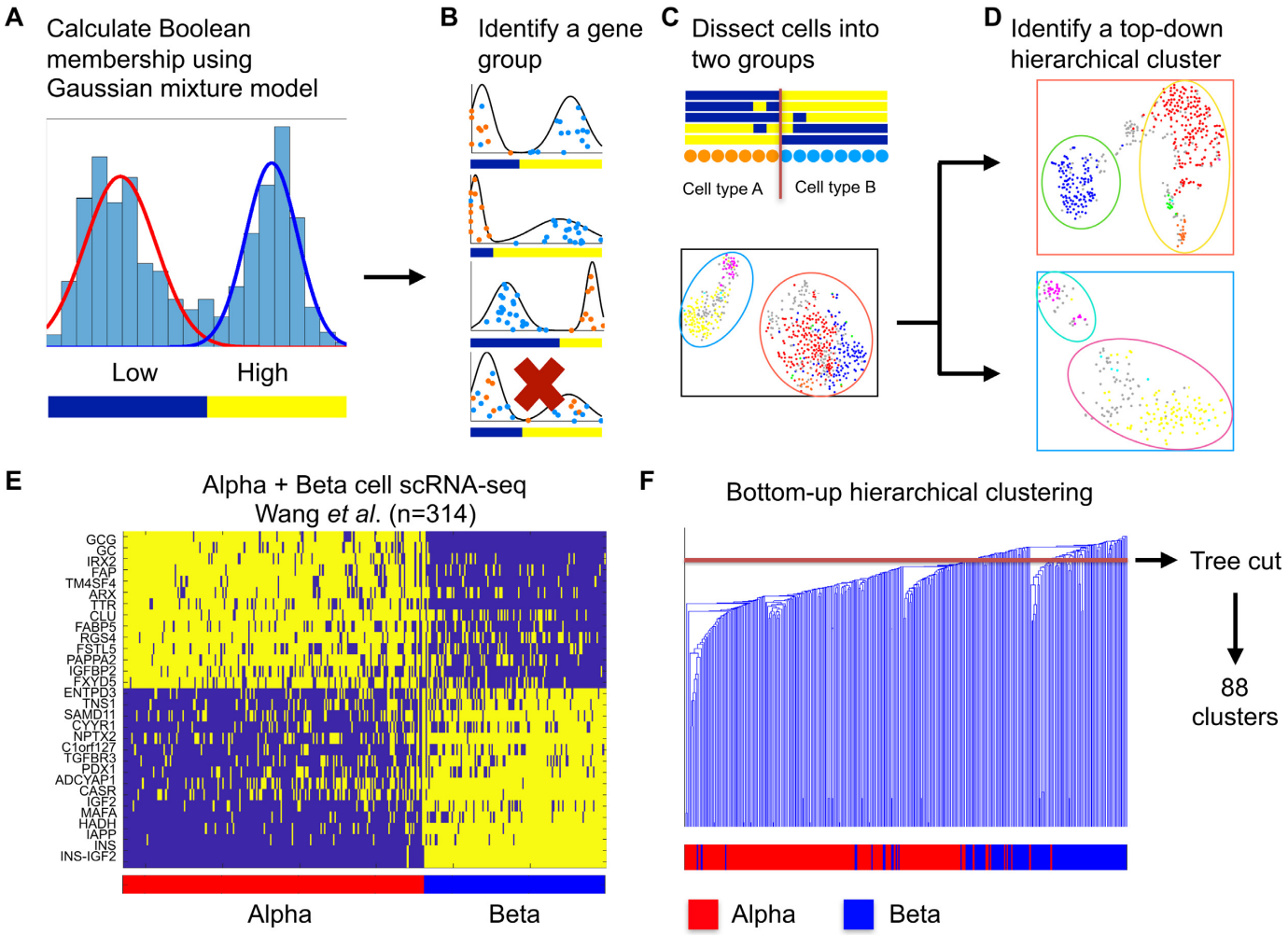

Figure 1. CellBIC implements top-down hierarchical clustering using bimodal pattern. (**A**) Step 1: Boolean membership is obtained using a Gaussian mixture model. (**B**) Step 2: A gene group is selected based on the Boolean membership. Only genes observed in one mode significantly are included. (**C**) Step 3: A membership matrix is obtained using the selected gene set. Cells are divided into two groups based on the membership matrix. (**D**) A top-down clustering is performed by applying A-C recursively. (**E**) A membership matrix obtained by CellBIC when using human pancreatic α and β cells (3). (**F**) A classical bottom-up hierarchical clustering using human pancreatic α and β cells (3). The point to cut the tree is not well defined for the bottom-up hierarchical clustering.

and third quadrants of the membership matrix, N denotes the number of the genes aligned to the two seed genes, and 0.7 is clustering score weight. The clustering score weight was determined after testing various weight values using the benchmarking data (Supplementary Figure S5). This algorithm is recursively applied to the subgroups until every dissection is discarded (Figure 1D) or the minimum number of cells in the cluster is less than a cutoff value. The CellBIC performance was robust to the change of the minimum number of cells especially with the cell number <50 (Supplementary Figure S6).

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis

All the GO (29) and KEGG (30) pathways analyses were implemented using Enrichr 2017 version (31).

RESULTS

A top-down clustering better dissects cells than the classical hierarchical clustering

To test the usefulness of using modality in clustering scRNA-seq data, we prepared a well-characterized cell groups composed of human pancreatic α and β cells (3). Applying CellBIC to this dataset, we identified Glucagon (*GCG*) and *INS* as the top pairing seed genes, which are the marker for α and β cells, respectively (Figure 1E). Along with *GCG*, we found classical α cell markers such as *ARX* (32) and *IRX2* (33) were aligned with *GCG* (Fisher's exact test P-value = 3.11e–28 and 1.33e–38, respectively). Also, β cell markers such as *PDX1* (34) and *MAFA* (35) were aligned well, as expected, with *INS* (Fisher's exact test P-value = 7.27e–28 and 5.45e–43, respectively).

The classical bottom up clustering approach identified pre-defined α and β cell groups as well. However, it did not provide a well-structured tree. When a tree is cut, the hi-

erarchical clustering approach did not clearly distinguish α against β cells, which could mislead the subsequent analysis (Figure 1F). Our test demonstrates that the top-down clustering identified functional sub-cell types and can better be applied to clustering scRNA-seq data.

CellBIC showed outstanding clustering performance in the comprehensive benchmarking tests

For comprehensive benchmarking, we compared the performance of CellBIC with various clustering approaches for scRNA-seq data including the classical hierarchical clustering and t-SNE (27) followed by DBSCAN (28). We also included recently developed SC3 (13) and SIMLR (14) for the comparison.

For this benchmarking we used scRNA-seq data for (i) 668 cells from human pancreas (3), (ii) 2544 cells from human pancreas (6), (iii) 3005 cells from mouse cortex (6) and (iv) 201 cells from mouse lung (5). We used the annotation of cells provided by each study as a gold standard. If a cell is not defined, we left it as undefined. Supplementary Table S1 summarizes the composition of cells that we used.

From the transcriptome of 668 human pancreas cells, CellBIC identified α, β, pancreatic polypeptide (PP), ductal, acinar, and mesenchymal cells successfully (Figure 2A). The classical bottom-up hierarchical clustering method successfully identified α, β, PP, ductal and mesenchymal cells. However, many α cells were misclassified. The t-SNE approach identified α, β, duct, acinar and mesenchymal cells, but it misclassified most of β cells. This is because of the difficulties in dissecting α against β cells that are located closely in the reduced dimension, where DBSCAN only selected the subset of β cells (Supplementary Figure S7). SC3 identified α and β cells well. However, SC3 failed to distinguish PP cells from α cells. SIMLR almost perfectly separated four types of endocrine cells, but it failed to distinguish ductal, acinar, and mesenchymal cells (Figure 2A).

With another 2544 cells from human pancreas, SIMLR showed similar identification performance to CellBIC (Figure 2B). However, both hierarchical clustering and t-SNE+DBSCAN failed in identifying α cells against β cells. SC3 missed or misclassified considerable proportions of α and β cells (Figure 2B). For mouse cortex cells, CellBIC, hierarchical clustering, t-SNE+DBSCAN, and SC3 showed comparable performance (Figure 2C). SIMLR, however, failed in distinguishing interneurons and two pyramidal cells (CA1 and S1). For mouse lung cells, only CellBIC and SIMLR distinguished E16 stage cells (Figure 2D) but SIMLR misclassified a considerable number of cells.

Figure 2E summarizes the overall performance using ARI for the datasets we used. Overall, CellBIC showed best or at least comparable results with other classifiers. CellBIC's performance was robust for all the tested datasets (ARI $>$ 0.7) SC3 showed comparable performance with CellBIC for 3 independent datasets. However, it performed poorly when using 2544 human pancreas cells. Our results demonstrate that clustering using modality may overcome the potential algorithmic biases in calculating distance between cells and/or subsequent analysis.

CellBIC reconstructs hierarchical cluster trees

We further investigated the hierarchical structure identified by CellBIC. From human pancreatic tissues (3), we first obtained two large cell groups. Gene ontology (GO) of the first group (left) showed terms related with endocrine function of pancreas such as 'Maturity onset diabetes' (adjusted *P*-value = 1.50e–6) and 'Insulin secretion' (adjusted *P*-value = 7.94e–6) (Figure 3 and Supplementary Table S2). The other group (right) showed terms related with exocrine function of pancreas such as 'Hippo signaling pathway' (adjusted *P*-value = 6.37e–3) (36,37) (Figure 3 and Supplementary Table S3). Endocrine cells were further divided into a group containing α (*GCG*+) and PP (*PPY*+) cells and the other group containing β (*INS*+) and δ (*SST*+) cells. This is consistent with the view that PP and δ cells are similar with α and β cells, respectively (38,39). The exocrine cells divided into mesenchymal cells expressed genes related with 'ECM-receptor interaction' (adjusted *P*-value = 1.73e–2) and a group expressed genes related with 'Tight junction' (adjusted *P*-value = 3.63e–4) (Figure 3 and Supplementary Tables S4–S5). The second group was further divided into ductal and acinar cells. In sum, CellBIC identified a structure of pancreatic cell types in a hierarchical form. We also obtained similar results from the scRNA-seq dataset for 2544 cells from human pancreatic tissues (6) (Supplementary Figure S8 and Supplementary Tables S6–S11).

CellBIC also reconstructed a hierarchical structure using the scRNA-seq data from mouse cortex (4). At the top level, mouse cortex cells were dissected into two large groups, one with a function related with 'Myelination' (adjusted *P*-value = 2.92e–2) and the other 'Synaptic transmission' (adjusted *P*-value = 2.62e–4), suggesting non-neuronal and neuronal cells, respectively (Supplementary Figure S9 and Supplementary Tables S12–S13). The neuronal cells were separated by the order of the interneurons and pyramidal neurons from two different sources including somatosensory (S1) and hippocampal CA1 region (Supplementary Figure S9 and Supplementary Tables S14–S15). The non-neuronal cells were separated by the order of oligodendrocytes, endothelial-mural cells, astrocyte-ependymal cells and microglia (Supplementary Figure S9 and Supplementary Tables S16–S19).

CellBIC identified the characteristics of β cells associated with T2D and aging

We further questioned the sub-clusters in the β cells composed of 98 adult and 20 child pancreatic β cells (3). CellBIC identified a group of cells with higher expression levels of *SIX3* and *CD14* (Figure 4A). While these cells showed strong *INS* expression levels, we found majority of cells high with *SIX3* and *CD14* were from T2D donors (Fisher's exact test *P*-value: 1.56e–14). To confirm this, we used an independent scRNA-seq data from normal and T2D β cells (18) (Figure 4B). In these independent datasets, genes highly expressed in *SIX3*+ cells (*CD14*, *NEFM*, *SIX3*, *ITGB3BP*) also had significantly higher expression in the cells from T2D donors (Wilcoxon rank sum test *P*-value $<$ 1.0e–2). Also, genes exclusively observed in *SIX3*+ cells such as *CPB1* and *CPA2* were expressed significantly less frequently

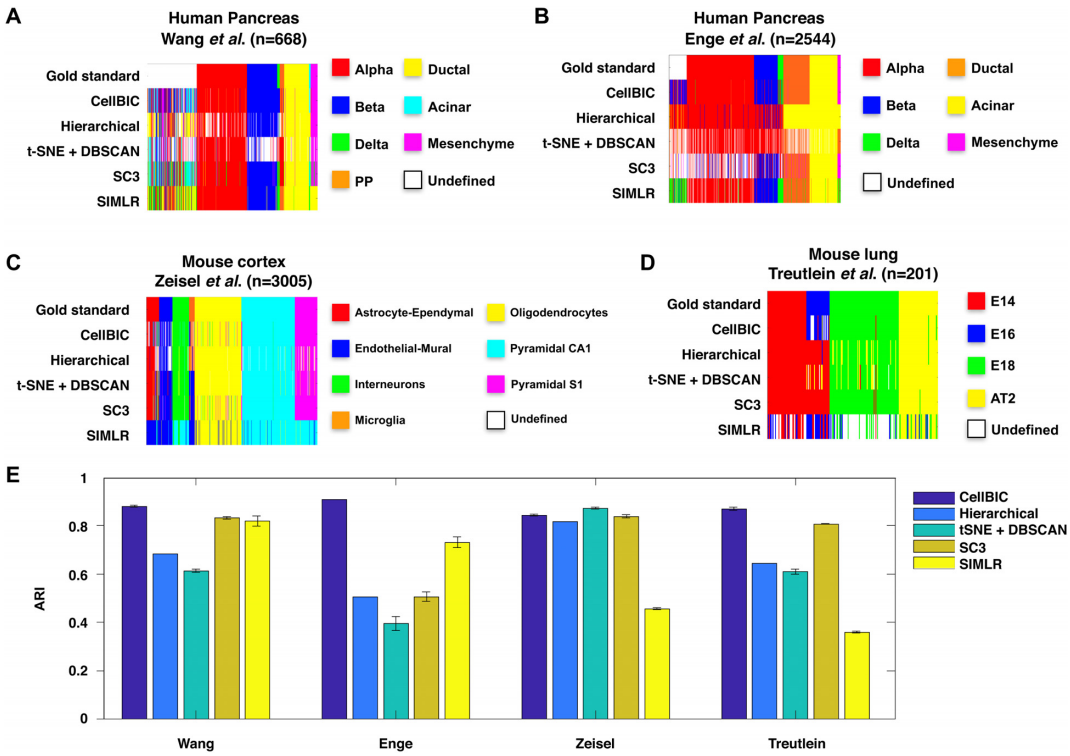

Figure 2. CellBIC outperforms in the benchmarking tests. The pre-defined sets (gold standard) are compared with the predicted cell types by CellBIC, bottom-up hierarchical clustering, t-SNE+DBSCAN clustering, SC3, and SIMLR. (**A**, **B**) two human pancreata (3,6), (**C**) a mouse cortex (4) and (**D**) a mouse liver dataset (5). (**E**) Performance evaluation of five clustering algorithms for four scRNA-seq datasets using ARI.

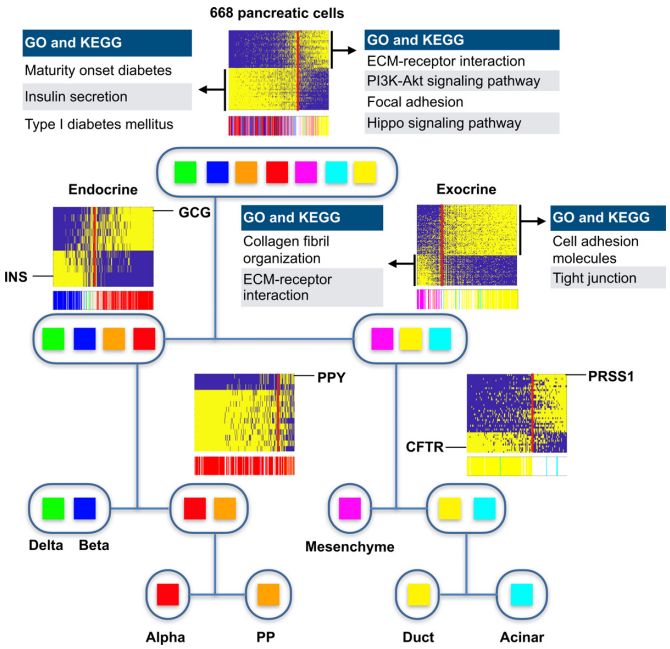

Figure 3. CellBIC reconstructs a top-down hierarchical structure from 668 human pancreatic cells (3). The entire human pancreatic cells were divided into endocrine and exocrine cells. Endocrine cells were further divided into a group containing α (*GCG*+) and PP (*PPY*+) cells and another group containing β (*INS*+) and δ (*SST*+) cells. The exocrine cells were divided into mesenchymal cells and a group consisting of ductal and acinar cells. Finally, CellBIC reconstructed the hierarchical structure for the pancreatic sub-types.

in normal β cells (Wilcoxon rank sum test *P*-value < 1.0e–4).

Interestingly, a previous genome-wide association study also showed a potential role of SIX3 in β cell maturation and a relevance of SIX3 with type 1 diabetes (T1D) and T2D risk (40). Furthermore, a subset of β cells expresses CD14, which is associated with the immune surveillance (41) and β cell viability (42). Our study may suggest CellBIC can be used to identify β cell markers associated with diabetes risk.

The clustering using the top second seed pair identified another β cell sub-type marked by N-cadherin (*CDH2*) (Figure 5A). We found *CDH2*+ cell group is populated with cells obtained from adult donors (Fisher's exact test *P*-value = 2.71e–16, Figure 5A). The age-dependent gene expressions of two seed genes *LCN2* and *CDH2* is also supported by a published bulk RNA-seq data obtained from human adult and fetal pancreatic β cells (43) (Wilcoxon rank sum test *P*-value < 0.01, Figure 5B). The validation using the two scRNA-seq datasets from donors with various ages (6,18) showed significantly positive correlations (correlation test *P*-value = 3.05e-4 and *P*-value = 3.12e–4, respectively, Figure 5C and D). Previous staining study identified the expression of *CDH2* in human β cells (44,45). We also found that *CDH2* is expressed only in adult β cells in the two independent scRNA-seq datasets (3,18) (Supplementary Figure S10). The negligible *CDH2* expression in adult α cells from the same donor further showed that *CDH2* expression in adult β cells are not from contamination such as doublet

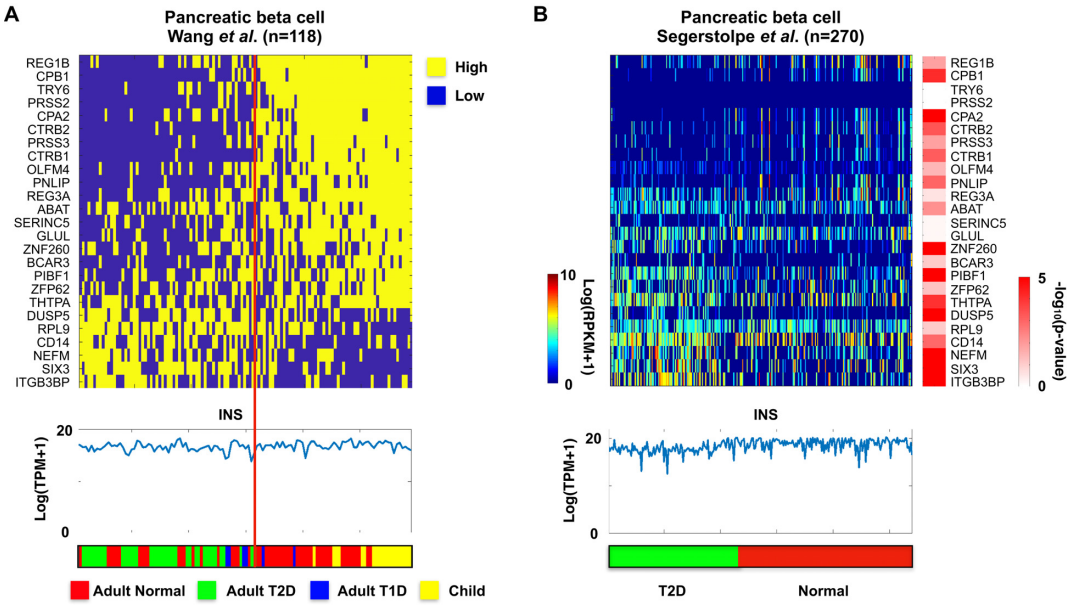

Figure 4. CellBIC identifies genes associated with T2D from β cell sub-clustering. (**A**) A membership matrix of sub-clustering of β cell shows *SIX3* and *CD14* are highly expressed in β cells from T2D donors. (**B**) Evaluation of the gene expression using the β cell scRNA-seq dataset from normal and T2D patients (18).

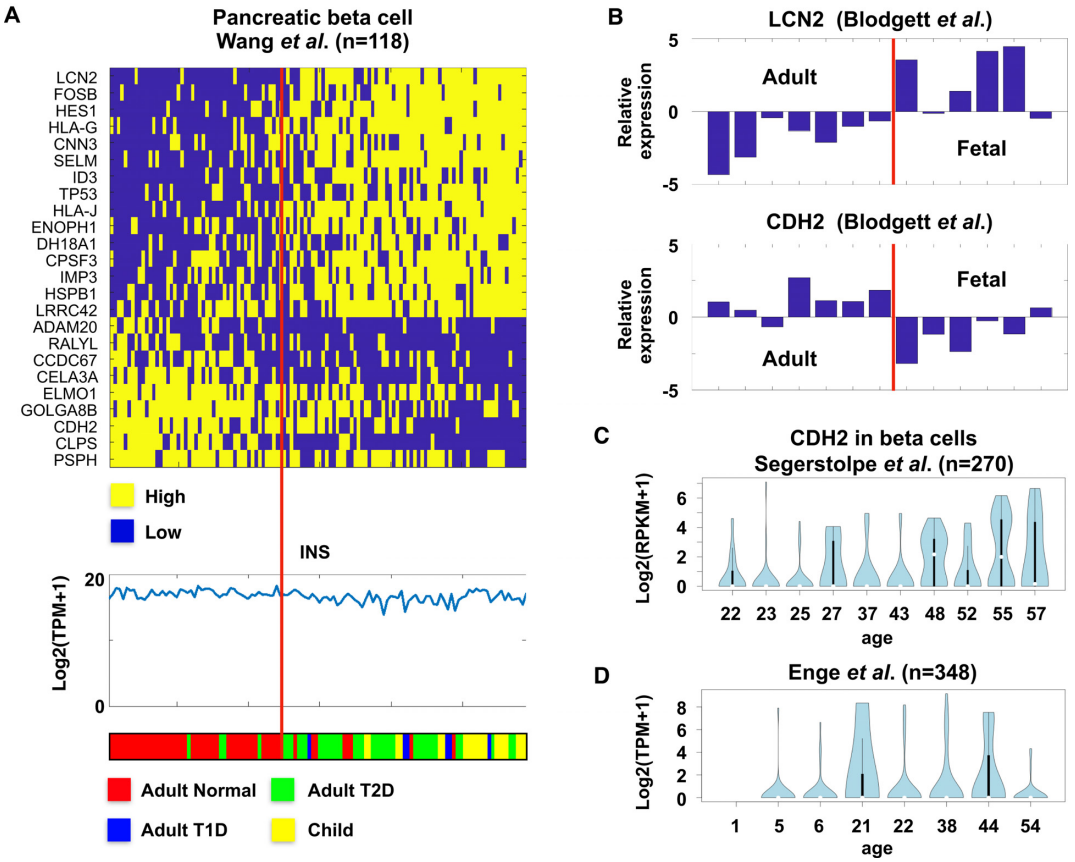

Figure 5. CellBIC identifies genes associated with aging from β cell sub-clustering. (**A**) A membership matrix of sub-clustering of β cell show *CDH2* is highly expressed in adult β cells. (**B**) *LCN2* and *CDH2* expression shows age-dependency in the bulk RNA-seq data obtained from fetal and adult pancreas (43). Violin plots show age-dependent *CDH2* expression in the two β cell scRNA-seq datasets by (**C**) Segerstolpe *et al.* (18) and (**D**) Enge *et al.* (6).

cells. Our study suggests that *CDH2* may be a β cell aging marker.

DISCUSSION

Clustering scRNA-seq data provides an unbiased way to characterize cells. CellBIC newly implemented a top-down hierarchical clustering approach for scRNA-seq data. To dissect cell groups, CellBIC designed a new flat clustering algorithm in a top-down clustering based on modality. Unlike its competitors, it does not require a calculation of cell-to-cell distance from the expression values directly. In the series of benchmarking tests, CellBIC outperformed the classical bottom-up hierarchical clustering approach. This may be because a simple distance metric the bottom-up clustering relies cannot deal with multi-dimensional data effectively. Ensemble approaches such as SIMLR and SC3 were designed to overcome potential biases using a single metric. Our benchmarking tests demonstrated that CellBIC outperformed the bottom-up hierarchical clustering approach as well as ensemble-based approaches. The robust performance of CellBIC in clustering four independent scRNA-seq datasets shows the advantages of using the distribution of data instead of using cell-to-cell distances.

Modality has been used for a number of studies using scRNA-seq data. MAST (46), SCDE (47), and scDD (48) utilized bimodal characteristics of gene expression in scRNA-seq to identify differentially expressed genes. These studies indicate that a gene with a bimodal expression pattern across single cells could be a good marker for distinguishing a cell state of cell type.

Previously, single cell based approaches identified *CFAP126* (also known as *Fltp*), *CD9*, and *ST8SIA1* as the β cell sub-type markers (49,50). We did not identify them in our analysis. However, CellBIC identified *SIX3* and *CDH2* as a marker for T2D and aging, respectively. These are obtained from the top 2 seed gene pairs. Interestingly, the other four algorithms could not identify these two β cell subtype markers using their default options (Supplementary Figures S11–S13). Interestingly, SC3 found a group of cells depleted with *SIX3* and *CDH2* expression when it was forced to identify more than three clusters (Supplementary Figure S12). Our results exhibit that there are algorithmic advantages of using the modality for cell clustering.

Even though we used bimodal pattern for our clustering, a multi-modal pattern will be observed if multiple cell types are observed in scRNA-seq. Multi-modal pattern can be identified using mixture of multiple Gaussian. CellBIC cannot currently use the multi-modal patterns. Clustering using multi-modal pattern requires large cell number for assigning membership to each mode and it may not be easy each to reconstruct the hierarchical structure. Instead, CellBIC identifies clusters by recursively identifying for bimodal patterns.

DATA AVAILABILITY

Input files for the benchmarking datasets and a MATLAB source code for CellBIC are available at https://github.com/neocaleb/CellBIC.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Klaus Kaestner for sharing his scRNA-seq data and critical review of the manuscript.
Author Contributions: K.J.W. conceived and designed the experiments. J.K. and D.S. performed the experiments and analyzed the data. J.K., K.J.W. and D.S. wrote the paper. The manuscript was approved by all authors.

FUNDING

American Diabetes Association grant [1-16-JDF-086 to D.E.S.]; NIH [R01DK-106027 to K.J.W.]; National Center for Advancing Translational Sciences [UL1TR001878]; Institute for Translational Medicine and Therapeutics (ITMAT) Transdisciplinary Program in Translational Medicine and Therapeutics (to K.J.W. and D.E.S.) (in part); Functional Genomics Core of the Penn Diabetes Research Center [to P30-DK19595]. Funding for open access charge: NIH [R01DK-106027].
Conflict of interest statement. None declared.

REFERENCES

1. Shapiro,E., Biezuner,T. and Linnarsson,S. (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.
2. Grün,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**, 251–255.
3. Wang,Y.J., Schug,J., Won,K.-J., Liu,C., Naji,A., Avrahami,D., Golson,M.L. and Kaestner,K.H. (2016) Single cell transcriptomics of the human endocrine pancreas. *Diabetes*, **65**, 3028–3038.
4. Zeisel,a., Manchado,a. B.M., Codeluppi,S., Lonnerberg,P., La Manno,G., Jureus,a., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.
5. Treutlein,B., Brownfield,D.G., Wu,A.R., Neff,N.F., Mantalas,G.L., Espinoza,F.H., Desai,T.J., Krasnow,M and Quake,S.R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, **509**, 371–375.
6. Enge,M., Arda,H.E., Mignardi,M., Beausang,J., Bottino,R., Kim,S.K. and Quake,S.R. (2017) Single-Cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, **171**, 321–330.
7. Jaitin,D.A., Kenigsberg,E., Keren-Shaul,H., Elefant,N., Paul,F., Zaretsky,I., Mildner,A., Cohen,N., Jung,S., Tanay,A. *et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, **343**, 776–779.
8. Xin,Y., Kim,J., Okamoto,H., Yancopoulos,G.D., Lin,C. and Correspondence,J.G. (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.*, **24**, 608–615.
9. Kolodziejczyk,A.A., Kim,J.K., Tsang,J.C.H., Ilicic,T., Henriksson,J., Natarajan,K.N., Tuck,A.C., Gao,X., Bühler,M., Liu,P. *et al.* (2015) Single cell RNA-Sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**, 471–485.
10. Patel,A.P., Tirosh,I., Trombetta,J.J., Shalek,A.K., Gillespie,S.M., Wakimoto,H., Cahill,D.P., Nahed,B. V., Curry,W.T., Martuza,R.L. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
11. Karypis,G., Han,E.-H. and Kumar,V. (1999) Chameleon: hierarchical clustering using dynamic modeling. *Computer*, **32**, 68–75.
12. Balcan,M.-F., Liang,Y. and Gupta,P. (2014) Robust hierarchical clustering. *J. Mach. Learn. Res.*, **15**, 4011–4051.

13. Kiselev,V.Y., Kirschner,K., Schaub,M.T., Andrews,T., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.
14. Wang,B., Zhu,J., Pierson,E., Ramazzotti,D. and Batzoglou,S. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414–416.
15. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
16. Villani,A.-C., Satija,R., Reynolds,G., Sarkizova,S., Shekhar,K., Fletcher,J., Griesbeck,M., Butler,A., Zheng,S., Lazo,S. *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.
17. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.
18. Segerstolpe,Å., Palasantza,A., Eliasson,P., Andersson,E.-M., Andréasson,A.-C., Sun,X., Picelli,S., Sabirsh,A., Clausen,M., Bjursell,M.K. *et al.* (2016) Single-Cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.*, **24**, 593–607.
19. Camp,J.G., Sekine,K., Gerber,T., Loeffler-Wirth,H., Binder,H., Gac,M., Kanton,S., Kageyama,J., Damm,G., Seehofer,D. *et al.* (2017) Multilineage communication regulates human liver bud development from pluripotency. *Nature*, **546**, 533–538.
20. Deng,Q., Ramsköld,D., Reinius,B., Sandberg,R., Ramskold,D., Reinius,B. and Sandberg,R. (2014) Single-Cell RNA-Seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
21. Langfelder,P., Zhang,B. and Horvath,S. (2008) Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R. *Bioinformatics*, **24**, 719–720.
22. Datta,S. and Datta,S. (2003) Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.
23. Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6745–6750.
24. Datta,S. and Datta,S. (2006) Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, **7**, S17.
25. Datta,S. and Datta,S. (2006) Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, **7**, 397.
26. Shalek,A.K., Satija,R., Adiconis,X., Gertner,R.S., Gaublomme,J.T., Raychowdhury,R., Schwartz,S., Yosef,N., Malboeuf,C., Lu,D. *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**, 236–240.
27. Van Der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **620**, 267–284.
28. Ester,M., Kriegel,H.P., Sander,J. and Xu,X. (1996) A Density-Based algorithm for discovering clusters in large spatial databases with noise. *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, **10**, 1.1.71.1980.
29. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
30. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
31. Kuleshov,M. V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.
32. Collombat,P., Mansouri,A., Hecksher-Sørensen,J., Serup,P., Krull,J., Gradwohl,G. and Gruss,P. (2003) Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes Dev.*, **17**, 2591–2603.
33. Petri,A., Ahnfelt-Rønne,J., Frederiksen,K.S., Edwards,D.G., Madsen,D., Serup,P., Fleckner,J. and Heller,R.S. (2006) The effect of neurogenin3 deficiency on pancreatic gene expression in embryonic mice. *J. Mol. Endocrinol.*, **37**, 301–316.
34. Kitamura,T., Nakae,J., Kitamura,Y., Kido,Y., Biggs,W.H., Wright,C.V.E., White,M.F., Arden,K.C. and Accili,D. (2002) The forkhead transcription factor Foxo1 links insulin signaling to Pdx1 regulation of pancreatic β cell growth. *J. Clin. Invest.*, **110**, 1839–1847.
35. Nishimura,W., Kondo,T., Salameh,T., El Khattabi,I., Dodge,R., Bonner-Weir,S. and Sharma,A. (2006) A switch from MafB to MafA expression accompanies differentiation to pancreatic β-cells. *Dev. Biol.*, **293**, 526–539.
36. Gao,T., Zhou,D., Yang,C., Singh,T., Penzo-Méndez,A., Maddipati,R., Tzatsos,A., Bardeesy,N., Avruch,J. and Stanger,B.Z. (2013) Hippo signaling regulates differentiation and maintenance in the exocrine pancreas. *Gastroenterology*, **144**, 1543–1553.
37. George,N.M., Day,C.E., Boerner,B.P., Johnson,R.L. and Sarvetnick,N.E. (2012) Hippo signaling regulates pancreas development through inactivation of yap. *Mol. Cell. Biol.*, **32**, 5116–5128.
38. DiGruccio,M.R., Mawla,A.M., Donaldson,C.J., Noguchi,G.M., Vaughan,J., Cowing-Zitron,C., van der Meulen,T. and Huising,M.O. (2016) Comprehensive alpha, beta and delta cell transcriptomes reveal that ghrelin selectively activates delta cells and promotes somatostatin release from pancreatic islets. *Mol. Metab.*, **5**, 449–458.
39. Lawlor,N., George,J., Bolisetty,M., Kursawe,R., Sun,L., Sivakamasundari,V., Kycia,I., Robson,P. and Stitzel,M.L. (2017) Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.*, **27**, 208–222.
40. Arda,H.E., Li,L., Tsai,J., Torre,E.A., Rosli,Y., Peiris,H., Spitale,R.C., Dai,C., Gu,X., Qu,K. *et al.* (2016) Age-dependent pancreatic gene regulation reveals mechanisms governing human β cell function. *Cell Metab.*, **23**, 909–920.
41. Osterbye,T., Funda,D.P., Fundová,P., Månsson,J.E., Tlaskalová-Hogenová,H. and Buschard,K. (2010) A subset of human pancreatic beta cells express functional CD14 receptors: A signaling pathway for beta cell-related glycolipids, sulfatide and β-galactosylceramide. *Diabetes. Metab. Res. Rev.*, **26**, 656–667.
42. Garay-Malpartida,H.M., Mourão,R.F., Mantovani,M., Santos,I.A., Sogayar,M.C. and Goldberg,A.C. (2011) Toll-like receptor 4 (TLR4) expression in human and murine pancreatic beta-cells affects cell viability and insulin homeostasis. *BMC Immunol.*, **12**, 18.
43. Blodgett,D.M., Nowosielska,A., Afik,S., Pechhold,S., Cura,A.J., Kennedy,N.J., Kim,S., Kucukural,A., Davis,R.J., Kent,S.C. *et al.* (2015) Novel observations from next-generation RNA sequencing of highly purified human adult and fetal islet cell subsets. *Diabetes*, **64**, 3172–3181.
44. Parnaud,G., Gonelle-Gispert,C., Morel,P., Giovannoni,L., Muller,Y.D., Meier,R., Borot,S., Berney,T. and Bosco,D. (2011) Cadherin engagement protects human β-cells from apoptosis. *Endocrinology*, **152**, 4601–4609.
45. Johansson,J.K., Voss,U., Kesavan,G., Kostetskii,I., Wierup,N., Radice,G.L. and Semb,H. (2010) N-cadherin is dispensable for pancreas development but required for β-cell granule turnover. *Genesis*, **48**, 374–381.
46. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
47. Kharchenko,P. V, Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.
48. Korthauer,K.D., Chu,L.-F., Newton,M.A., Li,Y., Thomson,J., Stewart,R. and Kendziorski,C. (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.*, **17**, 222.
49. Bader,E., Migliorini,A., Gegg,M., Moruzzi,N., Gerdes,J., Roscioni,S.S., Bakhti,M., Brandl,E., Irmler,M., Beckers,J. *et al.* (2016) Identification of proliferative and mature β-cells in the islets of Langerhans. *Nature*, **535**, 430–434.
50. Dorrell,C., Schug,J., Canaday,P.S., Russ,H.A., Tarlow,B.D., Grompe,M.T., Horton,T., Hebrok,M., Streeter,P.R., Kaestner,K.H. *et al.* (2016) Human islets contain four distinct subtypes of β cells. *Nat. Commun.*, **7**, 11756.