gCAnno: a graph-based single cell type annotation method

Xiaofei Yang[1,2¶], Shenghan Gao[2,3¶], Tingjie Wang[2,3¶], Boyu Yang[2,3], Ningxin Dang[4], Kai Ye[2,3,4,5*]

[1] School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2] MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[3] School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[4] Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, Shaanxi, China

[5] The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, China

* Corresponding author

E-mail: kaiye@xjtu.edu.cn (Kai Ye)

[¶]These authors contributed equally to this work.

Abstract

Background

Current single cell analysis methods annotate cell types at cluster-level rather than ideally at single cell level. Multiple exchangeable clustering methods and many tunable parameters have a substantial impact on the clustering outcome, often leading to incorrect cluster-level annotation or multiple runs of subsequent clustering steps. To address these limitations, methods based on well-annotated reference atlas has been proposed. However, these methods are currently not robust enough to handle datasets with different noise levels or from different platforms.

Results

Here, we present gCAnno, a graph-based Cell type Annotation method. First, gCAnno constructs cell type-gene bipartite graph and adopts graph embedding to obtain cell type specific genes. Then, naïve Bayes (gCAnno-Bayes) and SVM (gCAnno-SVM) classifiers are built for annotation. We compared the performance of gCAnno to other state-of-art methods on multiple single cell datasets, either with various noise levels or from different platforms. The results showed that gCAnno outperforms other state-of-art methods with higher accuracy and robustness.

Conclusions

gCAnno is a robust and accurate cell type annotation tool for single cell RNA analysis. The source code of gCAnno is publicly available at https://github.com/xjtu-omics/gCAnno.

Keywords: graph embedding, cell type annotation, single cell RNA analysis

Background

Bulk RNA sequencing measures average gene expression level in a large population of cells, hindering dissection of heterogeneous cell types [1]. In 2009, single cell RNA sequencing (scRNA-seq) technology was developed to provide valuable insights into cell heterogeneity [2].

In general, accurate cell type annotation for single cell data is a prerequisite for any further investigation of cell heterogeneous [3-6]. The commonly used cell type annotation methods, including Seurat [7], SCANPY [8] and SINCERA [9], adopts a similar procedure of data quality control, reads mapping, UMI quantification, expression normalization, clustering, differentially expressed genes (DEGs) of each cluster identification and cell type assignment based on biomarker genes [10]. However, those methods report cluster-level rather than truly single cell-level annotation results, masking subtle differences within each cluster. In addition, different clustering methods and many tunable parameters led to uncertain clustering outcome. These above two factors cause incorrect cluster-level annotations or multiple runs of subsequent clustering steps [10].

To overcome the above issues, two distinct strategies, namely biomarker-based and reference-based approaches, have been proposed. The biomarker-based methods, such as Garnett [11] and CellAssign [12], aim to establish mappings between the query dataset and the well-studied biomarkers. In particular, Garnett trains a classifier based on the user defined markup language. CellAssign builds a probabilistic model that leverages prior knowledge of cell-type marker genes for annotation. However, collecting a comprehensive biomarker set of different cell types is cumbersome, time-consuming and subjective [13]. Thus recently reference-based approaches, such as Scmap [14], Chetah [15] and scPred [16] have been developed and are gaining popularity after a number of well-annotated single cell data were published, especially the datasets released by human cell atlas (HCA) [17]. The reference-based methods follow data-driven strategy and construct mappings between query dataset and the well-annotated reference datasets. For example, Scmap uses drop-based method to select feature genes as variables and constructs mapping by distance and correlation coefficient. Another method, scPred selects differential principle components (PCs) calculated by gene expression value between cell types and trains an SVM model with these PCs. However, these methods are sensitive to experiment batches, sequencing platforms and noises, all of which are intrinsic properties of the single cell datasets.

Here, we propose a reference-based method, gCAnno, using graph representation feature selection strategy to comprehensively represent the global view of associations between cell types and genes for robust and high accuracy single cell-level annotation. Our gCAnno method starts with construction of a weighted cell type-gene bipartite graph. Then, graph embedding is applied to capture the cell type specific genes and naïve Bayes (gCAnno-Bayes) and SVM (gCAnno-SVM) classifiers are built for further annotation (Fig. 1). We compared gCAnno with the state-of-the-art methods on four published datasets as the basic test [3-6]. We also reported the performance comparison on large dataset with deep annotation level [18], different single cell platforms, simulated datasets with either various cell type imbalance situations and different dropout noise levels as the advanced test. Finally, runtime is summarized to demonstrate the efficiency of gCAnno.

Methods

Here we summarized the framework of gCAnno. gCAnno adopts graph structure for cell type specific gene set detection and accurate cell type annotation. Firstly, gCAnno builds cell type-gene bipartite graph based on gene expression abundances and intensities, in which gene expression abundance is the proportion of cells expressing the gene in a given cell type while intensity is the average expression in cells expressing the gene. Then, graph embedding is adopted to obtain the embedding vectors of gene nodes and cell type nodes. Next, gCAnno selects a set of genes for each cell type with similar profiles in the embedding space. Finally, based on the detected cell type specific genes, gCAnno trains naïve Bayes and SVM classifiers. The workflow of gCAnno is depicted in Fig. 1.

Cell type-gene bipartite graph construction

Starting from the well-annotated reference scRNA-seq data, we constructed a weighted cell type-gene bipartite graph (wCGBG) containing both cell type nodes (CTN) and gene nodes (GN). Edges between CTN and GN indicate the correlation of a gene and a cell type while weight W measures significance of correlation. The weight is calculated by:

$$w_{k,j} = \begin{cases} \dfrac{m_{k,j}}{n_k} \times mean\left(\overrightarrow{g_{k,j}}\right), & if \;\; n_k \neq 0 \\ 0 \qquad\qquad\qquad\qquad\quad , & others \end{cases} \tag{1}$$

where n_k is the cell count of cell type k, $m_{j,k}$ is the number of cells expressed gene j in cell type k. $\overrightarrow{g_{j,k}}$ is the expression vector of gene j in cell type k. W is the product of the gene expression abundance and intensity. We use gene expression abundance and intensity to establish a relationship between cell types and genes in the form of proportion to reduce the impact of individual gene loss (dropout) or cell number imbalance.

Graph embedding and cell type-gene specific relation detection

After wCGBG construction, we used node2vec to obtain the low dimensional vectors (the embedding vectors) of gene nodes and cell type nodes. The first step is construction of a neighborhood set $N(u)$ of each node u (either gene or cell type node) by a probability walk [19]. Then, we optimized the following objective function $f(u)$ by maximizing the log-probability of observing a neighborhood set.

$$\max_f \sum_{u \in V} \log P\left(N(u) \mid f(u)\right) \tag{2}$$

This optimization step enables the embedding vectors to capture the specificity and strength of interactions between cell node and gene node, e.g. if one gene is specific and highly expressed in one cell type, the corresponding two embedding vectors are similar. Then, we calculated Euclidean distance between the vector of genes and cell types. We selected top n (a user defined parameter, default $n = 65$, Additional file 1:Figure S1) closest genes for each cell type as the cell type specific gene set based on the overall performance on the five datasets we used [3-6][18].

Classifier construction

After obtaining the cell type specific gene set, we build naïve Bayes (gCAnno-Bayes) and SVM (gCAnno-SVM) classifiers for annotation. For gCAnno-SVM, we directly use the expression of cell type specific genes as features to train an SVM classifier. For gCAnno-Bayes, we build a binary matrix to presents cell type and its corresponding specific genes, e.g. the element $b_{ij} = 1$ indicates gene *j* is one of the specific genes in cell type *i*. We train a Bernoulli Naïve Bayes to get genes' conditional probability in each cell type and the prior probability of cell types. The query dataset is binarized and the annotation is based on maximum posterior probability of single cell's cell type specific genes expression.

Performance measurement and dataset

Performance assessment and comparison. Cell type annotation is a typical multi-classification problem. We applied kappa coefficient as the performance measurement of classification, defined as equation (3).

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = \frac{N_{corr}}{N_t}, \quad p_e = \frac{\sum_{i=1}^{K} a_i \times b_i}{N_t \times N_t} \qquad (3)$$

where N_{corr} is the ratio of total number of cells with corrected cell type annotation, N_{t} is the total number of cells in the dataset, K is the number of truly cell types, a_i is the number of corrected annotated cells in the *i*-th cell type, and b_i is the number of cells in the *i*-th cell type， p_o is the accuracy, $a_i \times b_i$ is the product of the actual and predicted quantity, p_e punishes bias for unbalance evaluation.

To evaluate the performance of gCAnno, we performed both cross-validation test and independent heterogeneous test (cross-platform test). First, we adopted the five-fold cross-validation strategy following recent single cell analysis comparison published earlier [13, 16] on four published datasets and simulated noise datasets to evaluate the overall and robustness performance (Additional file 2: File S1). Then, we performed independent test on datasets from different sequencing platforms (the cross-platform testing) to evaluate the generalization capability of gCAnno.

Tools in comparison. The calculation results of Scmap, Chetah and scPred were obtained from the corresponding publications [14-16]. For SVM, we followed the previous report [13] which is using drop-based method [20] for feature selection.

Datasets used in basic overall performance test. To illustrate the stable performance of gCAnno across various species and tissue types, we compared gCAnno with other methods using four published datasets, including liver, pancreas, Arabidopsis thaliana root (AT root), hepatocellular carcinoma and intrahepatic cholangiocarcinoma (HCC and ICCA) datasets (Table 1; Additional file 2: File S1; Additional file 3: Figure S2;

Additional file 4: Table S1). The true labels of the cells in each dataset are obtained from the corresponding publications.

Table 1 The list of scRNA-seq datasets in overall performance test.

Dataset	#Cells	#Genes	# Cell types
Liver [4]	8,444	20,007	14
Pancreas [3]	8,562	20,126	13
AT root [6]	7,053	32,833	19
HCC, ICCA [5]	4,729	19,379	8

Note: # means the number of

Large dataset with deep annotation level. To demonstrate the performance of gCAnno in large dataset (cell number more than 50,000) with deep annotation level (more than 20 cell types). We compared gCAnno with other methods in 20 mouse organs dataset with 54,246 cells, 29 cell types and 23,433 genes. The true labels of the cells in each dataset are also obtained from the original publications [18]. (Additional file 2: File S1; Additional file 5: Figure S3; Additional file 6: Table S2).

Simulated dropout and imbalance datasets. To evaluate the robustness of gCAnno in the presence of dropout noise, we simulated different dropout rates in four above datasets (Table 1), by modifying the expression level of a random gene subset (10%, 20%, 30%, 40% and 50% of all genes) to zero (Additional file 2: File S1). Similarly, we used five-fold cross validation to evaluate its performance. In each validation, we simulated the dropout noise in either training group (reference dropout) or test group (query dropout), and calculated the kappa coefficient for each method.

To simulate the cell number imbalance noise, we randomly sampled different proportions (0.1:1, 0.3:1, 0.5:1, 0.7:1, 0.9:1, 1:0.9, 1:0.7, 1:0.5, 1:0.3 and 1:0.1) of cell count in two cell types (Hepatocyte and GamaDetaT) in liver dataset as the reference data for classifier constructing. To get more accuracy testing, this simulation was repeated five times (Additional file 2: File S1).

Cross platform datasets. To compare cross platform performance (various studies using different sequencing platforms), we searched and identified four datasets suitable for this purpose, including two liver datasets from 10x and mCel-seq2 platforms and two pancreas datasets from drop-seq and smart-seq2 platforms (Table 2). We noticed that the cell type annotation labels of the same tissue from different platforms are not identical. Thus, we unified the labels by removing cell types absent in either of the datasets (Additional file 7: Figure S4; Additional file 8: Table S8; Additional file 2: File S1).

Table 2 The list of scRNA-seq datasets in cross platform test

Dataset	#Cells	#Genes	# Cell types	Platform
Liver [4]	8,103	20,007	7	10x
Pancreas [3]	8,037	20,126	9	Drop-seq

Liver [21]	7,130	33,941	7	mCel-seq2
Pancreas [22]	2,068	25,526	9	Smart-seq2

Note: # means the number of

Results

To evaluate the performance of gCAnno, we first evaluated the cell type-gene specific relation, and then compared gCAnno with five state-of-art methods, including Scmap-cell, Scmap-cluster, Chetah, scPred and SVM, in the following four aspects: 1) cell type specificity of gCAnno detected genes, 2) overall performance on different scRNA-seq datasets, 3) robustness test on simulated drop-out and imbalance noise data, 4) cross platform annotation.

Cell type specificity of gene sets detected by gCAnno

After graph embedding step, gCAnno selects cell type specific gene sets, which largely determines the performance of our approach. Thus, we first evaluated the cell type specificity of gene sets detected in the four datasets. We noticed that clear cell type specific expression patterns are observed for these selected genes (Additional file 9: Figure S5; Additional file 10: Figure S6; Additional file 11: Figure S7). Among the reported marker genes from the corresponding publications, gCAnno is able to capture an average of 57% of them, indicating gCAnno's effectiveness of cell type specific gene identification (Additional file 12: Figure S8; Additional file 13: Table S4).

Overall and large dataset performance evaluation

We next evaluated and compared overall performance of gCAnno, Scmap, scPred, Chetah and SVM with four published scRNA-seq datasets (Table 1). We found that the comprehensive kappa coefficient of both gCAnno was consistently much higher than those of Scmap-cluster, Scmap-cell and scPred, respectively ($p < 0.05$, Wilcoxon rank sum test) (Fig. 2a-2d) (Additional file 14: Table S5), hinting gCAnno's better performance than other methods on cell type annotation across different species (e.g. human or plant), organs (e.g. liver or pancreases), or disease states (e.g. health or cancer). In 20 mouse organs dataset, the comprehensive kappa coefficient of both gCAnno were 0.74 (gCAnno-Bayes) and 0.94 (gCAnno-SVM), and other methods achieve 0.16 (Scmap-cluster), 0.18 (Scmap-cell), 0.80 (Chetah), 0.63 (scPred) and 0.92 (SVM), respectively. We found that gCAnno-SVM achieved highest performance than other methods in large dataset with deep annotation level (Additional file 6: Table S2; Additional file 15: Figure S9).

Robustness on dropout and imbalance noisy data

Besides basic accuracy, we examined its robustness in the presence of different types of noises. Dropout and cell count imbalance noises are two major types and the most challenging in scRNA-seq data. Dropout is a technical noise in the form of missing value in gene expression [10], while cell number imbalance among cell types is coming from biology itself. We found gCAnno achieved the highest and rather stable kappa coefficients for both reference dropout and query dropout tests in four datasets (Fig. 3; Additional file 16: Figure S10; Additional file 17: Table S6; Additional file 18: Figure S11). Remarkably, gCAnno achieved average kappa coefficients of 0.88 (gCAnno-SVM) and 0.79 (gCAnno-Bayes) even when dropout rate was as high as 50%, while other methods achieve 0 (Scmap-cluster), 0.44 (Scmap-cell), 0.37 (Chetah), 0.25 (scPred) and 0.79 (SVM), respectively. Moreover, we found gCAnno, SVM and Scmap-cell achieved the highest and stable kappa coefficients (average values are about 0.99) for different cell count imbalance ratios (Additional file 16: Figure S10; Additional file 19: Table S7). All of these results show gCAnno is better than other methods for dropout and cell count imbalance noises and achieved the best performance on highly noisy data (e.g. 50% dropout rate and 1:0.1 imbalance rate), suggesting the effectiveness of the wCGBG in selecting accurate features in the presence of high noise.

Cross platform annotation

Different single cell sequencing platforms have platform specific features or bias [23], limiting cross platform cell type annotation. We evaluated the platform compatibility of gCAnno on two liver datasets and two pancreas datasets from four platforms (10x, mCel-seq2, Drop-seq, and Smart-seq2) (Table 2). We used one platform dataset as the training data and the other as the testing data. For the performance comparison, gCAnno achieved consistently high kappa coefficient values for liver dataset tests (Fig. 4a and Fig. 4b) and for pancreas dataset tests (Fig. 4c and Fig. 4d) (Additional file 20: Table S8). These results show gCAnno is able to maintain high annotation accuracy for real heterogeneous and cross platform data in the presence of systematic platform specific bias.

Runtime evaluation

Finally, we evaluated the runtime of gCAnno based on datasets in above tests (Additional file 21: Table S9; Additional file 22: Figure S12). We found that the time takes in model building (including graph construction and embedding) step is positive correlated with the number of graph nodes (Pearson's correlation is 0.94). Once the model has been built, the annotation step only takes less than 1 minute (e.g. for mCel-seq2 platform liver dataset with 8103 cells only takes 48 seconds).

Discussion

In this study, we present gCAnno, a novel graph-based cell type identification method for scRNA-seq data. The most significant feature of gCAnno is the construction of wCGBG, enabling gCAnno to capture the global characteristics of association between cell types and genes. This feature allows gCAnno to detect accurate feature genes for each cell type, leading to accurate annotation results and robustness for different noise types and rates. In addition, gCAnno is able to annotate not only human scRNA-seq, but also plant scRNA-seq (e.g. Arabidopsis data). Its stable and high performance across platforms, indicates wide application as a “pan-platform” method.

gCAnno contains SVM version (gCAnno-SVM) and naïve Bayes version (gCAnno-Bayes). The SVM version takes into account the effect of expression value while naïve Bayes version only considers the existence of cell type specific genes. From the evaluation result, the SVM version seems suitable for the dataset with deep annotation level and contains largely similar cell types between training and test sets. However, in cross platform datasets from different studies and different sequencing platforms, gene expression value might fluctuate significantly, rendering better performance of naïve Bayes version than SVM version.

Since gCAnno is a reference-based cell type annotation method, it lacks the ability to identify novel cell types. For novel type cells, gCAnno assigns the closest cell types with the most similar expression profiles to them, which might be reasonable in most of applications but probably require further improvement. Integrating the biomarker-based method for novel cell type annotation and reference-based method for accurate pre-defined cell type annotation, we think, will be one direction to explore.

Conclusion

We have implemented a stable and high-performance automated cell type annotation tool, gCAnno, for scRNA-seq datasets. With an easy use Python running script as an example, we hope gCAnno will be useful for the scRNA-seq data analysis.

Abbreviations

kappa: kappa coefficient

DEGs: differentially expressed genes

UMI: Unique Molecular Identifier

scRNA-seq: Single-cell RNA-seq

PCs: principle components

HCA: human cell atlas

wCGBG: weighted cell type-gene bipartite graph

10x: 10x Genomics platform

SVM: Support Vector Machine

Declarations

Ethics approval and consent to participate

This study used previously published data and did not obtain any new data directly involved humans, plants or animals.

Consent for publication

Not applicable.

Availability of data and materials

Datasets used for the analyses in this study are summarized in Additional file 2: File S1. The source code of gCAnno is publicly available at https://github.com/xjtu-omics/gCAnno.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Science Foundation of China (61702406 and 31671372), the National Science and Technology Major Project of China (grand no. 2018ZX10302205), and the National Key R&D Program of China (2018YFC0910400 and 2017YFC0907500), the General Financial Grant from the China Postdoctoral Science Foundation (2017M623178), and the Fundamental Research Funds for the Central Universities (xzy012020012).

Authors' contributions

KY and XY conceived the study. SG designed and performed the experiments. SG and BY analysed the data. SG developed the program. XY and SG wrote the manuscript. SG and ND completed figures of manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Peng Jia and Yongyong Kang for their important suggestions and feedback.

Authors' information

School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, China

Xiaofei Yang

MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, China

Kai Ye, Xiaofei Yang, Shenghan Gao, Tingjie Wang, Boyu Yang

School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi’an Jiaotong University, Xi’an, Shaanxi, China

Kai Ye, Shenghan Gao, Tingjie Wang, Boyu Yang

Genome Institute, the First Affiliated Hospital of Xi’an Jiaotong University, Xi’an, Shaanxi, China

Kai Ye, Ningxin Dang

The School of Life Science and Technology, Xi’an Jiaotong University, Xi’an, Shaanxi, China

Kai Ye

References

1. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA: **The technology and biology of single-cell RNA sequencing**. *Mol Cell* 2015, **58**(4):610-620.
2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A *et al*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**(5):377-382.
3. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM *et al*: **A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure**. *Cell Syst* 2016, **3**(4):346-360 e344.
4. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, Manuel J, Khuu N, Echeverri J, Linares I *et al*: **Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations**. *Nat Commun* 2018, **9**(1):4383.
5. Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, Rae Z, Hernandez JM, Davis JL, Martin SP *et al*: **Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer**. *Cancer Cell* 2019, **36**(4):418-430 e416.
6. Zhang TQ, Xu ZG, Shang GD, Wang JW: **A Single-Cell RNA Sequencing Profiles the Developmental Landscape of Arabidopsis Root**. *Mol Plant* 2019, **12**(5):648-660.
7. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell transcriptomic data across different conditions, technologies, and species**. *Nat Biotechnol* 2018, **36**(5):411-420.
8. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis**. *Genome Biol* 2018, **19**(1):15.
9. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y: **SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis**. *PLoS Comput Biol* 2015, **11**(11):e1004575.
10. Kiselev VY, Andrews TS, Hemberg M: **Challenges in unsupervised clustering of single-cell RNA-seq data**. *Nat Rev Genet* 2019, **20**(5):273-282.

11. Pliner HA, Shendure J, Trapnell C: **Supervised classification enables rapid annotation of cell atlases**. *Nat Methods* 2019, **16**(10):983-986.

12. Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, Wiens M, Walters P, Chan T, Hewitson B *et al*: **Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling**. *Nat Methods* 2019, **16**(10):1007-1015.

13. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A: **A comparison of automatic cell identification methods for single-cell RNA sequencing data**. *Genome Biol* 2019, **20**(1):194.

14. Kiselev VY, Yiu A, Hemberg M: **scmap: projection of single-cell RNA-seq data across data sets**. *Nat Methods* 2018, **15**(5):359-362.

15. de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP: **CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing**. *Nucleic Acids Res* 2019, **47**(16):e95.

16. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE: **scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data**. *Genome Biol* 2019, **20**(1):264.

17. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M *et al*: **The Human Cell Atlas**. *Elife* 2017, **6**.

18. Tabula Muris C, Overall c, Logistical c, Organ c, processing, Library p, sequencing, Computational data a, Cell type a, Writing g *et al*: **Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris**. *Nature* 2018, **562**(7727):367-372.

19. Grover A, Leskovec J: **node2vec: Scalable Feature Learning for Networks**. *KDD* 2016, **2016**:855-864.

20. Andrews TS, Hemberg M: **M3Drop: dropout-based feature selection for scRNASeq**. *Bioinformatics* 2019, **35**(16):2865-2867.

21. Aizarani N, Saviano A, Sagar, Mailly L, Durand S, Herman JS, Pessaux P, Baumert TF, Grun D: **A human liver cell atlas reveals heterogeneity and epithelial progenitors**. *Nature* 2019,

572(7768):199-204.

22. Segerstolpe A, Palasantza A, Eliasson P, Andersson EM, Andreasson AC, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK *et al*: **Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes**. *Cell Metab* 2016, **24**(4):593-607.

23. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W: **Comparative Analysis of Single-Cell RNA Sequencing Methods**. *Mol Cell* 2017, **65**(4):631-643 e634.

Figure legends

Fig 1. Overview of gCAnno. (a) Cell type-gene graph building. The graph contains gene nodes (gray circles) and cell type nodes (other color circles). **(b)** Graph embedding converts graphs into low dimensional vectors. Genes are selected based on the distance between the two types of vectors. **(c)** Training Naïve Bayes and SVM classifiers for annotation. **(d)** Cell type annotation for new query dataset.

Fig. 2. Overall performance evaluation. Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred, Chetah and SVM based on kappa coefficient on **(a)** liver dataset, **(b)** pancreas dataset, **(c)** HCC & ICCA dataset, and **(d)** AT root dataset. *: *p*-values < 0.1; **: *p*-values < 0.05; ***: *p*-values < 0.01, Wilcoxon rank sum test. The number is the mean of five cross validation. The error bar is the standard deviation. The *y*-axis is the kappa coefficient.

Fig. 3. Robustness performance evaluation. Robustness of dropout noise comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred, Chetah and SVM on **(a)** liver reference dropout dataset, **(b)** liver query dropout dataset, **(c)** pancreas reference dropout dataset, **(d)** pancreas query dropout dataset. The middle point is the mean kappa coefficients of five-fold cross validation. The error bar is the standard deviation. The *y*-axis is the kappa coefficient and the *x*-axis is the dropout rate.

Fig 4. Platform compatibility evaluation. Performance comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred, Chetah and SVM on cross platform datasets. (**a**) liver datasets, where reference is mCel-seq2 and query is 10x; (**b**) liver datasets, where reference is 10x and query is mCel-seq2; (**c**) pancreas dataset, where reference is drop-seq and query is smart-seq2 (**d**) pancreas datasets, where reference is smart-seq2 and query is drop-seq. The reference is the training data and the query is the testing data.

Additional files

Additional file 1.pdf: Figure S1. The test of gCAnno parameter top closest genes in five evaluation datasets. The parameter is stable in 25 to 85. When top gene select less than 5 (in all datasets) and more than 125 (in Arabidopsis and liver datasets), the performance are not well. In our evaluation, the default top closest genes in each cell type is 65 and user can adjustment by themselves.

Additional file 2.docx: File S1. Supplementary Materials, including data preparation, cell type information of each datasets, and supplementary methods.

Additional file 3.pdf: Figure S2. The tSNE plot of (a) liver, (b) pancreas, (c) HCC & ICCA and (d) AT root datasets.

Additional file 4.xlsx: Table S1. The tSNE result, cell barcodes and cell type labels of (a) liver, (b) pancreas, (c) HCC & ICCA and (d) AT root datasets.

Additional file 5.pdf: Figure S3. The tSNE plot of a large dataset with deep annotation level (20 mouse organs).

Additional file 6.pdf: Table S2. The large dataset kappa coefficient result (Fig 2e) and tSNE result.

Additional file 7.pdf: Figure S4. The tSNE plot of (a) mCel-seq2 liver, (b) 10x liver, (c) Drop-seq pancreas and (d) Smart-seq2 pancreas.

Additional file 8.xlsx: Table S3. The tSNE result, cell barcodes and cell type labels of (a) mCel-seq2 liver, (b) 10x liver, (c) Drop-seq pancreas and (d) Smart-seq2 pancreas.

Additional file 9.pdf: Figure S5. The heatmap of each cell type specific genes expression in four datasets (top closest gene number 65). It shows an obvious pattern in diagonal.

Additional file 10.pdf: Figure S6. The tSNE of embedding vectors of cell type nodes and gene nodes in four datasets. The selecting gene nodes are in red color and non-selecting gene nodes are in grey. The cell type nodes are blue triangles.

Additional file 11.pdf: Figure S7. An example of top 2 specific genes in each cell type of liver dataset. In tSNE plot, each gene specific expressed in corresponding cell type in red color. The shade of color means the expression value.

Additional file 12.pdf: Figure S8. The overlap of reported marker genes from the corresponding publications in four datasets with selected genes. The circle is selected genes and the square is not selected genes. The marker genes have different color and non-marker genes are gray.

Additional file 13.xlsx: Table S4. The statistic of select state of reported marker genes from the corresponding publications in four datasets with selected genes.

Additional file 14.xlsx: Table S5. The statistic of kappa coefficient in overall performance test (Fig 2a-d).

Additional file 15.pdf: Figure S9. The heatmap of each cell type specific genes expression in large dataset (top closest gene number 65). It shows an obvious pattern in diagonal.

Additional file 16.pdf: Figure S10. Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred, Chetah and SVM on (a) HCC and ICCA reference dropout dataset, (b) HCC and ICCA query dropout dataset, (c) AT root reference dropout dataset, (d) AT root query dropout dataset and (e) imbalance dataset.

Additional file 17.xlsx: Table S6. The statistic of kappa coefficient in dropout test.

Additional file 18.pdf: Figure S11. An example of the existence of selected cell type specific genes in liver ref dropout test dataset. The red color in more than one type means these types shared this gene. With the increasing of dropout rate, the degree of shared specific genes increased a little, but the specific pattern is still strong even in dropout rate 50%.

Additional file 19.xlsx: Table S7. The statistic of kappa coefficient in imbalance test.

Additional file 20.xlsx: Table S8. The statistic of kappa coefficient in cross platform test.

Additional file 21.xlsx: Table S9. Runtime statistic of gCAnno.

Additional file 22.pdf: Figure S12. The plot of building model time and graph scale. The building model time is correlated with graph node number (correlation coefficient is 0.94).