

1 **gCAnno: a graph-based single cell type annotation method**

2 Xiaofei Yang^{1,2¶}, Shenghan Gao^{2,3¶}, Tingjie Wang^{2,3¶}, Boyu Yang^{2,3}, Ningxin Dang⁴,

3 Kai Ye^{2,3,4,5*}

4 ¹ School of Computer Science and Technology, Faculty of Electronic and Information

5 Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

6 ² MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic

7 and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

8 ³ School of Automation Science and Engineering, Faculty of Electronic and

9 Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

10 ⁴ Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an,

11 Shaanxi, China

12 ⁵ The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an,

13 Shaanxi, China

14 * Corresponding author

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

16 ¶These authors contributed equally to this work.

17

18 **Abstract**

19 **Background**

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

27 **Results**

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

35 **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 a naïve Bayes classifier is 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 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.

Materials and 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, gCAnno trains a naïve Bayes classifier based on a binary matrix of cell types and the detected cell type specific genes. 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:

$$w_{k,j} = \begin{cases} \frac{m_{k,j}}{n_k} \times \text{mean}(\overrightarrow{g_{k,j}}), & \text{if } n_k \neq 0 \\ 0, & \text{others} \end{cases} \quad (1)$$

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 [18] 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 [18]. Then, we optimized the following objective function $f(u)$ by maximizing the log-probability of observing a neighborhood set.

$$\max_f \sum_{u \in V} \log P(N(u) | f(u)) \quad (2)$$

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 65 closest genes for each cell type as the cell type specific gene set based on the overall performance on the four datasets we used [3-6] (Additional file 1:Figure S1).

Classifier training

After obtaining the cell type specific gene set, we build a binary matrix of cell types and the corresponding gene set to train a Naïve Bayes classifier to get genes' conditional probability in each cell type and the prior probability of cell types. The query dataset 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 [19] as the performance measurement of classification, defined as equation (3).

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

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.

To evaluate the performance of gCAnno, we performed both cross-validation test and independent test. First, we adopted the five-fold cross-validation strategy following recent single cell analysis comparison published earlier [13, 16] on four published dataset and its simulated noise datasets to evaluate the basic overall and

robustness performance (Additional file 2: File S1). Then, we performed independent test on datasets from different sequencing platform (the cross platform testing) to evaluate the generalization capability of gCAnno.

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 3: Figure S2; Additional file 2: File 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

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, 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 4: Figure S3; Additional file 5: Table S1; 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 [20]	7,130	33,941	7	mCel-seq2
Pancreas [21]	2,068	25,526	9	Smart-seq2

174 Note: # means the number of

175 **Results**

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

182 **Cell type specificity of gene sets detected by gCAnno**

183 After graph embedding step, gCAnno selects cell type specific gene sets detected,
184 which largely determines the performance of our approach. Thus, we first evaluated
185 the cell type specificity of gene sets detected in the four datasets. We noticed that
186 clear cell type specific expression patterns are observed for these selected genes
187 (Additional file 6: Figure S4; Additional file 7: Figure S5). Among the reported
188 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 8: Figure S6; Additional file 9: Table S2).

Basic overall performance evaluation

We next evaluated and compared overall performance of gCAnno, Scmap, ScPred and Chetah with four published scRNA-seq datasets (Table 1). We found that the comprehensive kappa coefficient of gCAnno was consistently much higher than those of Scmap-cluster, Scmap-cell and scPred, respectively ($p < 0.05$, Wilcoxon rank sum test) (Fig. 2) (Additional file 10: Figure S7; Additional file 11: Table S3), 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).

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 12: Figure S8). Remarkably, gCAnno achieved average kappa coefficients of 0.78 even when dropout rate was as high as 50%, while other methods achieve 0.002 (Scmap-cluster), 0.41 (Scmap-cell), 0.37 (Chetah), and 0.29 (ScPred), respectively (Additional file 13: Table S4). Moreover, we found gCAnno and Scmap-cell achieved the highest and stable kappa coefficients (average values are about 0.99) for different cell count imbalance ratios (Additional file 12: Figure S8; Additional file 14: Table S5). 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 [22], 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). 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 15: Figure S9; Additional file 16: Table S6). These results show gCAnno is able to maintain high annotation accuracy for real 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 17: Table S7; Additional file 18: Figure S10). 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

243 high performance across platforms, indicates wide application as a “pan-platform”
244 method.

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

252 **Conclusion**

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

256 **Abbreviations**

257 **kappa:** kappa coefficient

258 **DEGs:** differentially expressed genes

259 **UMI:** Unique Molecular Identifier

260 **scRNA-seq:** Single-cell RNA-seq

261 **PCs:** principle components

262 **HCA:** human cell atlas

263 **wCGBG:** weighted cell type-gene bipartite graph

264 **10x:** 10x Genomics platform

265 **Declarations**

266 **Ethics approval and consent to participate**

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

269 **Consent for publication**

270 Not applicable.

271 **Availability of data and materials**

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

275 **Competing interests**

276 The authors declare that they have no competing interests.

277 **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) and the General Financial Grant from the China Postdoctoral Science Foundation (2017M623178).

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

300 **Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China**
 301 Kai Ye, Shenghan Gao, Tingjie Wang, Boyu Yang
 302 **Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University,**
 303 **Xi'an, Shaanxi, China**
 304 Kai Ye, Ningxin Dang
 305 **The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an,**
 306 **Shaanxi, China**
 307 Kai Ye

308 **References**

- 309 1. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA: **The**
 310 **technology and biology of single-cell RNA sequencing.** *Mol Cell* 2015,
 311 **58(4):610-620.**
- 312 2. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J,
 313 Tuch BB, Siddiqui A *et al*: **mRNA-Seq whole-transcriptome analysis of a**
 314 **single cell.** *Nat Methods* 2009, **6(5):377-382.**
- 315 3. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH,
 316 Wagner BK, Shen-Orr SS, Klein AM *et al*: **A Single-Cell Transcriptomic**
 317 **Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell**
 318 **Population Structure.** *Cell Syst* 2016, **3(4):346-360 e344.**
- 319 4. MacParland SA, Liu JC, Ma XZ, Innes BT, Bartczak AM, Gage BK, Manuel J,
 320 Khuu N, Echeverri J, Linares I *et al*: **Single cell RNA sequencing of human**
 321 **liver reveals distinct intrahepatic macrophage populations.** *Nat Commun*
 322 2018, **9(1):4383.**
- 323 5. Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, Rae Z, Hernandez
 324 JM, Davis JL, Martin SP *et al*: **Tumor Cell Biodiversity Drives**
 325 **Microenvironmental Reprogramming in Liver Cancer.** *Cancer Cell* 2019,
 326 **36(4):418-430 e416.**
- 327 6. Zhang TQ, Xu ZG, Shang GD, Wang JW: **A Single-Cell RNA Sequencing**
 328 **Profiles the Developmental Landscape of Arabidopsis Root.** *Mol Plant*
 329 2019, **12(5):648-660.**
- 330 7. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R: **Integrating single-cell**
 331 **transcriptomic data across different conditions, technologies, and species.**
 332 *Nat Biotechnol* 2018, **36(5):411-420.**

- 333 8. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene**
334 **expression data analysis.** *Genome Biol* 2018, **19**(1):15.
- 335 9. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y: **SINCERA: A Pipeline for**
336 **Single-Cell RNA-Seq Profiling Analysis.** *PLoS Comput Biol* 2015,
337 **11**(11):e1004575.
- 338 10. Kiselev VY, Andrews TS, Hemberg M: **Challenges in unsupervised**
339 **clustering of single-cell RNA-seq data.** *Nat Rev Genet* 2019, **20**(5):273-282.
- 340 11. Pliner HA, Shendure J, Trapnell C: **Supervised classification enables rapid**
341 **annotation of cell atlases.** *Nat Methods* 2019, **16**(10):983-986.
- 342 12. Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A,
343 Wiens M, Walters P, Chan T, Hewitson B *et al*: **Probabilistic cell-type**
344 **assignment of single-cell RNA-seq for tumor microenvironment profiling.**
345 *Nat Methods* 2019, **16**(10):1007-1015.
- 346 13. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT,
347 Mahfouz A: **A comparison of automatic cell identification methods for**
348 **single-cell RNA sequencing data.** *Genome Biol* 2019, **20**(1):194.
- 349 14. Kiselev VY, Yiu A, Hemberg M: **scmap: projection of single-cell RNA-seq**
350 **data across data sets.** *Nat Methods* 2018, **15**(5):359-362.
- 351 15. de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP: **CHETAH:**
352 **a selective, hierarchical cell type identification method for single-cell RNA**
353 **sequencing.** *Nucleic Acids Res* 2019, **47**(16):e95.
- 354 16. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE: **scPred:**
355 **accurate supervised method for cell-type classification from single-cell**
356 **RNA-seq data.** *Genome Biol* 2019, **20**(1):264.
- 357 17. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E,
358 Bodenmiller B, Campbell P, Carninci P, Clatworthy M *et al*: **The Human Cell**
359 **Atlas.** *Elife* 2017, **6**.
- 360 18. Grover A, Leskovec J: **node2vec: Scalable Feature Learning for Networks.**
361 *KDD* 2016, **2016**:855-864.
- 362 19. Cohen J: **Weighted kappa: nominal scale agreement with provision for**
363 **scaled disagreement or partial credit.** *Psychol Bull* 1968, **70**(4):213-220.
- 364 20. Aizarani N, Saviano A, Sagar, Mailly L, Durand S, Herman JS, Pessaux P,
365 Baumert TF, Grun D: **A human liver cell atlas reveals heterogeneity and**
366 **epithelial progenitors.** *Nature* 2019, **572**(7768):199-204.
- 367 21. Segerstolpe A, Palasantza A, Eliasson P, Andersson EM, Andreasson AC, Sun
368 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.

22. 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 a Naïve Bayes classifier for annotation. (d) Cell type annotation for new query dataset.

Fig. 2. Overall performance evaluation. Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred and Chetah on (a) liver dataset, (b) pancreas dataset, (c) HCC & ICCA dataset, and (d) AT root dataset. **: p-values < 0.05; ***: p-values < 0.01, Wilcoxon rank sum test.

Fig. 3. Robustness performance evaluation.

Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred and Chetah on (a) liver reference dropout dataset, (b) liver query dropout dataset, (c) pancreas reference dropout dataset, (d) pancreas query dropout dataset.

Fig 4. Platform compatibility evaluation. Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred and Chetah on liver datasets, (a) reference is mCel-seq2 and query is 10x, (b) reference is 10x and query is mCel-seq2; Comparisons of gCAnno with Scmap-Cluster, Scmap-Cell, scPred and Chetah on pancreas datasets, (c) reference is drop-seq and query is smart-seq2, (d) reference is smart-seq2 and query is drop-seq.

Additional files

Additional file 1.pdf: Figure S1. The test of the parameter.

Additional file 2.docx: File S1. Supplementary Materials.

Additional file 3.pdf: Figure S2. tSNE results for Datasets used in overall performance test.

Additional file 4.pdf: Figure S3. tSNE results for Datasets used for cross-platform test.

Additional file 5.xlsx: Table S1. tSNE results of all datasets we used.

Additional file 6.pdf: Figure S4. Expression patterns of detected cell type specific gene sets.

Additional file 7.pdf: Figure S5. The tSNE of embedding vector.

406 **Additional file 8.pdf: Figure S6. The tSNE of embedding vector with marker**
407 **genes.**

408 **Additional file 9.xlsx: Table S2. The marker genes for each cell type.**

409 **Additional file 10.pdf: Figure S7. The snakey diagram of four published datasets**
410 **used in overall performance test.**

411 **Additional file 11.xlsx: Table S3. The result of overall performance evaluation.**

412 **Additional file 12.pdf: Figure S8. The result of dropout test in AT root and**
413 **HCC&ICCA datasets and imbalance test.**

414 **Additional file 13.xlsx: Table S4. The result of dropout test.**

415 **Additional file 14.xlsx: Table S5. The result of imbalance test.**

416 **Additional file 15.pdf: Figure S9. The snakey diagram of datasets used in**
417 **cross-platform test.**

418 **Additional file 16.xlsx: Table S6. The result of platform compatibility evaluation.**

419 **Additional file 17.xlsx: Table S7. Runtime of gCAnno.**

420 **Additional file 18.pdf: Figure S10. Runtime of building model step.**