SCSA: a cell type annotation tool for single-cell RNA-seq data

Yinghao Cao^, Xiaoyue Wang*, Gongxin Peng*

Center for Bioinformatics, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Beijing, 100005, China

^ first author

* corresponding author

Yinghao Cao (yhcao@ibms.pumc.edu.cn)

Xiaoyue Wang (wxy@ ibms.pumc.edu.cn)

Gongxin Peng (penggongxin@ibms.pumc.edu.cn)

Abstract:

Currently most methods take manual strategies to annotate cell types after clustering the single-cell RNA-seq data. Such methods are labor-intensive and heavily rely on user expertise, which may lead to inconsistent results. We present SCSA, an automatic tool to annotate cell types from single-cell RNA-seq data, based on a score annotation model combining differentially expressed genes and confidence levels of cell markers in databases. Evaluation on real scRNA-seq datasets from different sources shows that SCSA is able to assign the cells into the correct types at a fully automated mode with a desirable precision.

Introduction

Recent development of single-cell RNA sequencing (scRNA-seq) methods has enabled unbiased, high-resolution transcriptomic analysis of individual cells in a heterogeneous cell population [1-4]. scRNA-seq methods have been used to characterize thousands to millions of cells in developing embryos [5], immune cells [6], complex tissues such as brain [7] and tumor [8], which have greatly promoted our understanding of human development and diseases.

At the core of myriad scRNA-seq applications is the ability to identify different cell types and cellular states from a complex cell mixture based on gene expression profiles. Unsupervised clustering methods such as principal components analysis (PCA) and T-distributed stochastic neighbor embedding (t-SNE) have been developed to partition the cells based on the similarity of their gene expression patterns [9, 10]. Although computational toolkits such as Cell Ranger[11] and Seurat [12] have automated the analysis steps from raw data processing to cell clustering, they leave it to the user to provide the biological interpretation of these cell clusters [10].

A common practice is to manually assign a cell type to each cluster based on differentially expressed genes between clusters, by consulting the literature for cell-type specific gene markers. However, it is not only a labor-intensive process, but also may generate biased results with uncontrolled vocabularies for cell type labels,

making it impossible to compare between different datasets. Expert-curated knowledge databases such as CellMarker [13] and CancerSEA [14], have been developed to provide a comprehensive and unified resource of cell markers for various cell types in human and mouse tissues. Yet it is still rely on the users to manually label the cells based on the information. An automated tool is needed for the reproducibility and consistency of cell type annotation.

To overcome these difficulties and to streamline the cell type assignment process for scRNA-seq data, we developed SCSA, an algorithm that can automatically assign cell types for each cell cluster in scRNA-seq data. SCSA follows the logic of the manual annotation that marker genes of known cell types highly expressed in a cell cluster could be used to label the cluster. To mimic the human decision-making process, SCSA exploits a score annotation model that accounts for differentially expressed genes, the confidence levels of the marker genes and discrepancies of marker genes in different cell marker databases. For cell clusters lacking known cell markers, SCSA will also perform a GO enrichment analysis and report the results to give some clues to the user. We evaluated the performance of SCSA on several real single cell datasets, which generated from different platforms including Smart-seq2 [4] and 10x Genomics[11]. We demonstrated that SCSA successfully classified the cell types in these datasets.

Materials and methods

Marker genes identification

The input of SCSA is a gene expression matrix, in a format that is supported by the output of CellRanger or Seurat. Based on these cells-genes expression matrix, SCSA identifies the marker genes of each cell cluster through differential gene expression analysis with log2-based fold-change (LFC) value and P-value (LFC >= 1, P <= 0.05). For each cell cluster, a marker gene identification vector is generated for j genes with LFC values, which is defined as $E_j = \{e_1, e_2, \cdots e_j\} = (e)_{j\times 1}$, here, e represents the absolute value of LFC multiplied by mean of all differentially

expressed genes.

Cell marker database

In order to improve the accuracy of cell cluster annotation for single-cell RNA-seq, SCSA uses cell markers from two public databases: CellMarker and CancerSEA. Up to now, CellMarker database contains 13,605 manually curated cell markers of 467 cell types in 158 human tissues or sub-tissues, and CancerSEA database provides a cancer single-cell functional state atlas, involving 14 functional states of 41,900 cancer single cells from 25 cancer types. Furthermore, SCSA can accept users-based marker gene database as additional information for cell cluster annotation. The users-based marker gene database must have two columns, with the first column as the name of a cell type and the second column as a marker gene name of the cell type. In that case, SCSA will combine both known databases and the custom database to predict the annotations for cell clusters.

Annotation model construction

For those genes which existed in both the DEGs and known cell marker databases, SCSA constructs a cell-gene sparse matrix (defined as $M_1 = (a_{ij})_{c_1 \times g_1}$) with $c_1 (c_1 \leq \text{i})$ cells and $g_1 (g_1 \leq \text{j})$ genes as "marker evidence". Here, for each cell i and each gene j in the matrix M_1, a refers to the sum number of references in the CellMarker database. To eliminate the huge differences of marker evidence between the well-known gene and less-known genes, we transform the value to log2-based and plus a constant (0.05). Also, to represent the whole gene set for a certain cell, we create a cell style vector which takes multiplication of standard deviation of the marker evidence and marker numbers (defined as $L_1 = \{l_1, l_2, \cdots l_{c1}\}$, where $l = std(a_{ij}) * num(a_{ij} > 0)$). So, for the known marker database, the raw score vector of a cell type is $S_{c_1} = M_1 \times E_{g1} * L_1 = \{s_1, s_2, \cdots s_{c_1}\}$.

For marker databases from multiple sources including known and users-defined, we define k as the total number of databases. Then k cell-gene sparse matrix could

be constructed according to the DEGs, which are defined as $M_k = (a_{ij})_{p\times q},\ p \times q \in \{(c_1 \times g_1), (c_2 \times g_2), \cdots, (c_k \times g_k)\}$. We define $E_k = (e)_{q\times 1}, q \in \{g_1, g_2, \cdots, g_k\}, (q \leq \mathrm{j})$ as multiple gene expression vectors. Then k raw score vectors will be generated, which is defined as $S_k = M_k E_k L_k = (a_{ij})_{p\times q} \times (e)_{q\times 1} * (l)_{q\times 1} = (s)_{p\times 1}, p \in \{c_1, c_2, \cdots c_k\}$. Suppose S_k to be a function to the score vectors, which is defined as F_l. To eliminate the difference of those vectors and compare them with each other standardly, SCSA performs z-score normalization for them. In detail,

$$F_l = M_l E_l L_l = (f_l)_{p\times 1}, l \in [1, k]$$

$$Z_l = (z_{mn})_{p\times 1} = (f'_l)_{p\times 1}, l \in [1, k]$$

$$f'_l = \frac{f_l - \bar{f}}{d}, \bar{f} = \frac{1}{p}\sum_{l=1}^{p} f_l, d = \sqrt{\frac{1}{p-1}\sum_{l=1}^{p}(f_l - \bar{f})^2}$$

Notably, the score vectors derived from different kinds of databases may have different lengths. To give a uniform score to a certain cell type, SCSA transforms them to the same length:

$$Z'_l = (z_{m'n'})_{p'\times 1}, z_{m'n'} = \begin{cases} z_{mn}, & m = m', n = n' \\ 0, & m \neq m' \parallel n \neq n' \end{cases}, p' = num\{c_1 \cup c_2 \cup \cdots \cup c_k\}$$

Finally, an annotation model will be constructed by merging the database weight coefficient matrix W and the last uniform score vector.

$$S' = (Z'_1, Z'_2, \cdots, Z'_k)W + b, W = (w_z)_{k\times 1}$$

GO enrichment analysis

Not all gene markers of cell types have been curated in known cell marker databases. To solve this problem, SCSA employs a GO enrichment analysis to allow identification of new cell types. In detail, for a certain GO term, SCSA uses Fisher's exact test to calculate the P-values, using the differentially expressed genes of the selected cluster as foreground and DEGs in other clusters and background values, respectively. After that, P-value was adjusted by the Benjamini-Hochberg (BH) method [15].

Real datasets

To assess the performance of SCSA, we used two kinds of real datasets. The first is the scRNA-seq data of experimentally mixed cells with known cell types. It includes four real datasets (GSE72056 [16], GSE81861 [17], E-MTAB-6149 [18], E-MTAB-6653 [18] and Fantom5 [19]), which were downloaded from NCBI GEO database. The other is the real tissue sample dataset with unknown cell types. It includes six peripheral blood mononuclear cells (PBMCs) datasets from a healthy donor of *Homo sapiens* downloaded from the 10X Genomics official website (*https://www.10xgenomics.com/)*.

Evaluation of SCSA performance

For the known cell type datasets, we defined cell type cluster by their real cell types. Then we generated calculated the differentially expression genes for each clusters using Student's t-test. We obtained the final list of differentially expression genes through setting the threshold with P-value of 0.05 and LFC value of 1.5. Finally, we compared the cell types predicted by SCSA with the real cell type to check the accuracy.

For the PBMC datasets, data preprocessing, normalization and unsupervised clustering were already performed by CellRanger workflow from its website. Also 10X Genomics official website illustrated a workflow example using 3k peripheral blood mononuclear cells (PBMCs) from a healthy donor containing 5 cell clusters and gave a final annotation results, they were monocytes, T cells, NK cells, megakaryocytes, and B cells, respectively. Here, we termed them as the "reference cell type annotation". In order to compare SCSA with CellRanger and Seurat uniformly for cell type annotation, we downloaded six datasets from 10X Genomics official website (*https://www.10xgenomics.com/*). Then, all six datasets were first processed using the CellRanger software (cellranger count). Notably, we diSCSArded the results of 1k and 10k PBMCs datasets because their existing clusters cannot be annotated without enough differentially expressed genes. So, for the rest four PBMCs datasets, we compared these results by SCSA with the "reference cell type annotation" to assess the performance of SCSA.

SCSA identified the marker genes of each cluster through the LFC (LFC>=1.5) value and P-value (P<=0.05). And then SCSA calculates the score vector of verified cell types that contain these marker genes based on these databases. In the annotation score model of SCSA, the cell type having the highest score in the score vector of verified cell types was used as the final annotation result for a cluster.

To evaluate the stability of SCSA in annotating the cell type of a cluster, we calculated the percentage of the five cell types (monocytes cells, T cells, NK cells, megakaryocytes cells, and B cells) in the four PBMCs datasets, respectively. And we generated the abundance of the same cell type in the four PBMCs datasets. To further demonstrate the performance of SCSA, we clustered all cell types of top five scores in a cell cluster as a heat map from the four PBMCs datasets using hierarchical clustering method.

Software availability

SCSA is implemented in Python3 as an open source software under the GNU General Public License, and the source code is freely available together with full documentation at https://github.com/bioinfo-ibms-pumc/SCSA.

Results

The SCSA algorithm is a three-step procedure that includes marker genes identification, annotation model construction, and gene ontology (GO) enrichment analysis (Figure 1). First, the input of SCSA is a gene expression matrix with cell cluster information (such as the output results of CellRanger or Seurat). SCSA identifies a group of marker genes for each cluster from input expression matrix by differential gene expression analysis. Next, for each cluster, genes identified as marker genes that have one or more linked cell types in a database will be used to generate a cell-gene matrix for that cluster. For each cell type in the matrix, SCSA then used a decision model to assign a score by combining the enrichment of marker gene expression and the strength of evidence for the marker genes in the database. SCSA could also take marker gene information from multiple databases and assign

different weights to them. Finally, if none of the marker genes of a certain cluster exists in known databases, SCSA performs an alternative gene ontology enrichment analysis step to annotate that cell cluster.

We first benchmarked the performance of SCSA on four scRNA-seq datasets from mixed cells with known cell types (Table 1). The GSE72056 dataset have five cell types that have been annotated manually by experts. Compared with it, SCSA predicted four cell types precisely except for “Cancer-associated fibroblast (CAF)”. For the CAF cell cluster, SCSA gave a different label named “Mesenchymal stem cell”. We checked the evidence in the CellMarker database and found that the “Mesenchymal stem cell” type had 34 marker genes but CAF only had 17 marker genes . Similarly, for seven known clusters in GSE81 861 datasets, SCSA correctly predicted six clusters except for the “Fibroblast” group. SCSA also gave a different annotation named “Mesenchymal stem cell” with 19 evidence marker genes instead of “Fibroblast” with 11 marker genes. This phenomenon was also found in public datasets [18]. It was reported that fibroblasts shared more common features with mesenchymal stem cells [20] by expressing similar cell immunophenotypic markers, as well as the genes that are known to be expressed in stem cells [21]. So we considered that “Mesenchymal stem cell” may also be suitable for the disputable cluster.

In addition, we also tested SCSA on the FANTOM5 dataset, which contains both human and mouse cell data. SCSA achieved a 52.4% and a 52% accuracy for human and mouse data respectively, when we removed the cell clusters containing lower than 10 cells. Interestingly, if we filtered the clusters containing lower than 40 cells, the accuracy of SCSA were improved to 75% and 72% for human and mouse data, respectively. To test whether the cell number of each cluster may have an impact on the accuracy of SCSA, we compared the cell number between the clusters with positive and negative prediction. As illustrated in Figure 2, the cell number with positive predictions was significantly different with that of negative predictions. So we demonstrated that more cells in clusters will improve the accuracy of SCSA. For

cell clusters with low cells, SCSA was likely to give a random cell type prediction, which may be due to the lack of consensus marker genes from the database for such clusters.

In order to evaluate the robustness of SCSA, we used four PBMCs (3k, 4k, 6k and 8k) datasets from 10X genomics website. We collected all possible cell types of a cell cluster according to the top five scores under the score annotation model of SCSA. The correlation of all cell types and scores were calculated and compared. As shown in Figure 3A, based on the five annotated cell types (monocyte cells, T cells, NK cells, megakaryocytes cells, and B cells) of CellRanger, SCSA achieved a great consistency in the four PBMCs datasets. Notably, the “macrophage cell” type was predicted as the second top score in a cluster, which was annotated as “monocytes” by SCSA due to the reason that they share many marker genes.

To further demonstrate the robustness of SCSA over the five annotated cell types (monocytes cells, T cells, NK cells, megakaryocytes cells, and B cells), we compared their abundance in each cluster using the four PBMCs datasets. As shown in Figure 3B, the percentages of cell numbers for five cell types annotated by SCSA remained stable across these datasets. T cells occupied half of the PBMCs, and monocytes cell represented another 25%, B cells and NK cells had similar levels, while megakaryocytes cell has the lowest number among all the five cell types. These results were consistent with the reference information of PBMCs.

Discussion

Currently, for scRNA-seq data, cell type annotation of cell clusters after unsupervised clustering is mainly conducted manually. The limitation of the manual procedure makes it impossible to generate high-quality, reproducible, and standardized annotation results for the growing number of scRNA-seq datasets. To address this challenge, we presented a novel tool, SCSA, for automatic annotating the cell types from single-cell RNA sequencing data, which can be applied directly on the output generated from CellRanger or Seurat. By introducing the newly developed

annotation model merging DEG and cell markers reference information to replace the manual steps, SCSA can perform the annotation task at a high accuracy and efficient level.

In cell type annotation, it is usually hard to find high-quality marker genes to describe a cell cluster. A strategy is to use genes specifically expressed in a cell cluster to mark the cell type. However, using a few marker genes is often not sufficient to distinguish a cell cluster from the others. In addition, using the whole expressed gene sets may decrease the power to find the true patterns within each cell cluster. Therefore, we used differentially expressed genes (DEGs) in the marker gene identification step in SCSA. This step avoids the influence of ubiquitously expressed genes and collects the appropriate genes for calculating the optimal score in the annotation model.

There still exist some limitations, which may influence the accuracy of cell type annotation using SCSA. First, the quantity of marker genes in these cell marker databases greatly impacted the results of cell type annotation. Since cell marker collection is far away from completion, it is possible that some cell types are unclassifiable due to the lack of appropriate markers. Specifically, this phenomenon is quite common for unknown tissues and novel sub-clusters of cells at different states. User-defined marker combinations need to be developed to solve this problem. SCSA can accept them as additional information to improve the annotation results. Second, for complex tissues such as cancer tissues, the accuracy of cell annotation is heavily relied on the clustering algorithms. Different unsupervised clustering method could have different results, especially when the cluster size is unevenly distributed in the population [10]. In that situation, algorithms using supervised clustering may be more appropriate for cell type classification [22].

Compared with the results of SCSA over different datasets, SCSA exhibited a reasonable accuracy and robustness in cell type annotation. Further efforts could be made to improve the annotation ability of SCSA by taking into account more information (e.g., the more accurate information of cell marker genes, the comprehensive clustering algorithm). We believe that SCSA is an important addition

to the toolbox used for single-cell studies and will greatly improve our efficiency and capacity to explore the functional potential of novel cell types.

Reference

1. Tang, F.C., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nature Methods, 2009. **6**(5): p. 377-U86.
2. Haque, A., et al., *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.* Genome Medicine, 2017. **9**.
3. Kolodziejczyk, A.A., et al., *The technology and biology of single-cell RNA sequencing.* Mol Cell, 2015. **58**(4): p. 610-20.
4. Picelli, S., et al., *Smart-seq2 for sensitive full-length transcriptome profiling in single cells.* Nature Methods, 2013. **10**(11): p. 1096-1098.
5. Chu, L.F., et al., *Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm.* Genome Biology, 2016. **17**.
6. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-240.
7. Zhong, S.J., et al., *A single-cell RNA-seq survey of the developmental landSCSApe of the human prefrontal cortex.* Nature, 2018. **555**(7697): p. 524-+.
8. Chung, W., et al., *Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer.* Nature Communications, 2017. **8**.
9. Bacher, R. and C. Kendziorski, *Design and computational analysis of single-cell RNA-sequencing experiments.* Genome Biology, 2016. **17**.
10. Kiselev, V.Y., T.S. Andrews, and M. Hemberg, *Challenges in unsupervised clustering of single-cell RNA-seq data (vol 20, pg 273, 2019).* Nature Reviews Genetics, 2019. **20**(5): p. 310-310.
11. Zheng, G.X.Y., et al., *Massively parallel digital transcriptional profiling of*

single cells. Nature Communications, 2017. **8**.

12. Butler, A., et al., *Integrating single-cell transcriptomic data across different conditions, technologies, and species.* Nature Biotechnology, 2018. **36**(5): p. 411-+.
13. Zhang, X.X., et al., *CellMarker: a manually curated resource of cell markers in human and mouse.* Nucleic Acids Research, 2019. **47**(D1): p. D721-D728.
14. Yuan, H.T., et al., *CancerSEA: a cancer single-cell state atlas.* Nucleic Acids Research, 2019. **47**(D1): p. D900-D908.
15. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Statistical Methodology, 1995. **57**(1): p. 289-300.
16. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-196.
17. Li, H.P., et al., *Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors (vol 50, pg 1754, 2018).* Nature Genetics, 2018. **50**(12): p. 1754-1754.
18. Lambrechts, D., et al., *Phenotype molding of stromal cells in the lung tumor microenvironment.* Nature Medicine, 2018. **24**(8): p. 1277-+.
19. Lizio, M., et al., *Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals.* Nucleic Acids Research, 2017. **45**(D1): p. D737-D743.
20. Haniffa, M.A., et al., *Mesenchymal stem cells: the fibroblasts' new clothes?* Haematologica-the Hematology Journal, 2009. **94**(2): p. 258-263.
21. Brohem, C.A., et al., *Comparison between fibroblasts and mesenchymal stem cells derived from dermal and adipose tissue.* International Journal of Cosmetic Science, 2013. **35**(5): p. 448-457.
22. Pliner, H.A., J. Shendure, and C. Trapnell, *Supervised classification enables rapid annotation of cell atlases.* Nature Methods, 2019. **16**(10): p. 983-+.

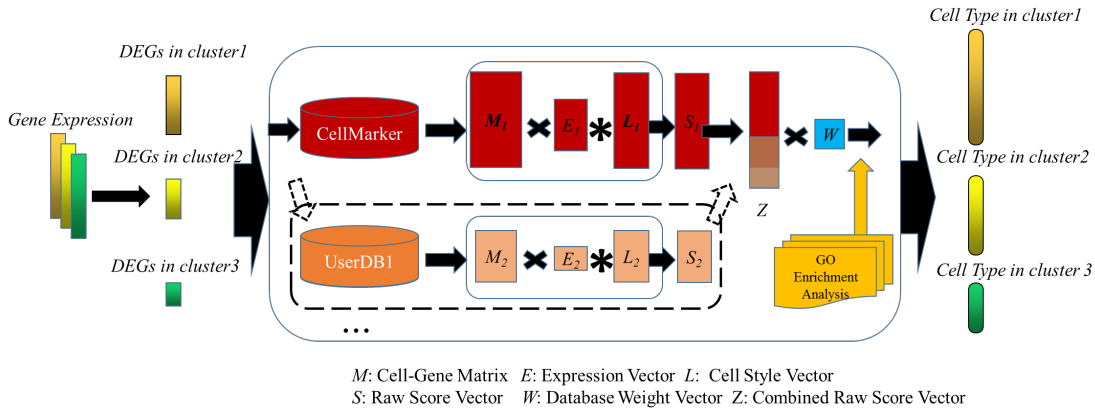


Figure 1. Flowchart of the SCA

First, DEGs of each cluster will be extracted and filtered from gene expression file. Next, SCSA employs marker gene databases to annotate cell clusters. In this step, both known marker gene database and user-defined marker database could be used simultaneously. For each cluster each database, a cell-gene matrix (*M*) with two vectors (*E*, *L)* will be generated to form a raw score vector (*S*). If multiple databases were selected, vectors would be normalized and combined together to make a new vector (*Z*), then multiplied with a database weight matrix (*W*) to make the last uniform vector. In the last step, ranked cell type vector will be generated according to the uniform score. In addition, SCSA employs GO enrichment analysis to give users some clue for unidentified clusters.

DataSet	Cell Number	Number of clusters	Accuracy	Species
GSE72056	2,840	5	100%*	Human
GSE81861	266	7	100%*	Human
Lambrechts D, et al (E-MTAB-6149 and E-MTAB-6653)	45,251	7	100%*	Human
Fantom5(>10)	1,280	24	54%	Human
Fantom5(>10)	1,329	20	55%	Mouse
Fantom5(>40)	996	11	73%	Human
Fantom5(>40)	1,048	8	75%	Mouse

Table 1. The predicted results of four known datasets

Four known datasets were selected to test the accuracy of SCA. For GSE72056, GSE81861, and *Lambrechts D, et al* (E-MTAB-6149 and E-MTAB-6653), SCSA successfully predicted all clusters precisely except for the "Cancer-associated fibroblast" cluster in GSE72056, "Fibroblast" cluster in GSE81861 and *Lambrechts D, et al* (E-MTAB-6149 and E-MTAB-6653). For the three clusters, SCA gave the same cell type "Mesenchymal stem cell" instead. Since more marker genes were found in the predicted cell type than the raw cell type and the cell types were similar and associated with each other, we thought SCA gained 100% accuracy prediction for the three datasets subjectively. For the first two Fantom5 datasets, cluster with cell number larger than 10 for each species were retained. While in the last two datasets, the threshold of cell number in cluster was up to 40.

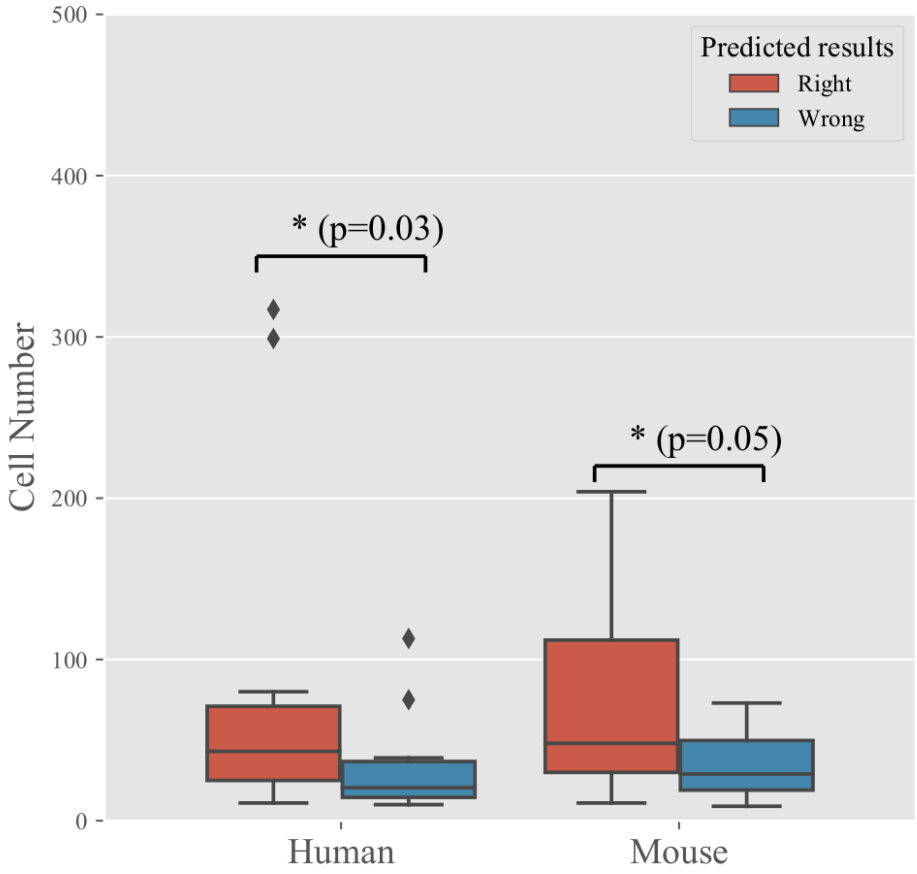
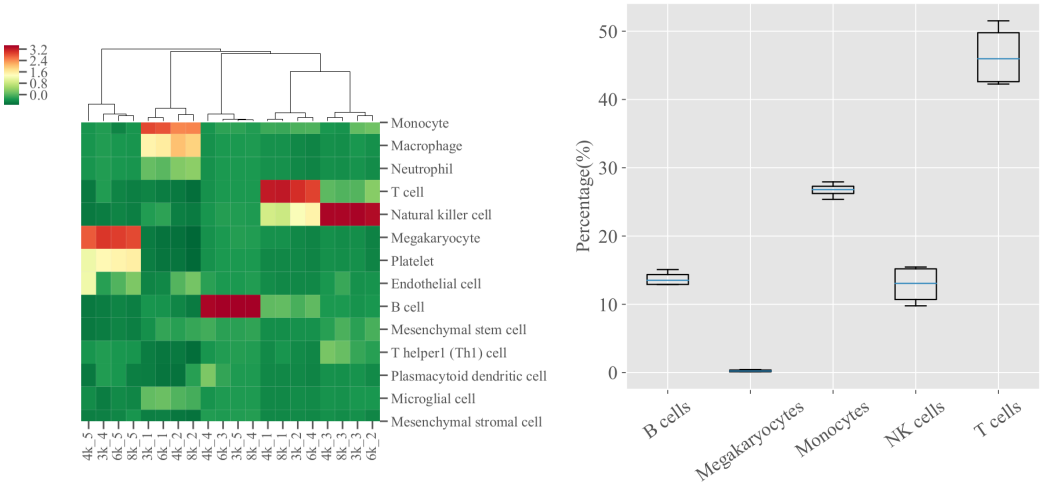


Figure 2. The comparison of cell numbers in cluster for FANTOM5 datasets prediction. Datasets were split according to the species. P-values were calculated using the Student's t-test,



A B

Figure 3. Cell components of PBMC cells predicted by SCSA. A. Clustering of uniform scores of top 5 predicted cell types in 4 PBMC sets by SCA. Each column stands for one cluster of 4 PBMC sets and each row stands for one cell type. Uniform scores were normalized using z-score method to make clusters comparable. B. Steady percentage of different cells in 4 PBMC sets. Percentage of cell number of each predicted cell type in 4 PBMC sets was calculated.