Unifying single-cell annotations based on the Cell Ontology

Sheng Wang[1,2,*], Angela Oliveira Pisco[3,*], Jim Karkanias[3], Russ B. Altman[1,2,3,#]

[1]*Department of Bioengineering, Stanford University, Stanford, CA 94305, USA.*
[2]*Department of Genetics, Stanford University, Stanford, CA 94305, USA.*
[3]*Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*

[*]*These authors contributed equally to this work*

[#]*Email:russ.altman@stanford.edu*

Abstract

Single cell technologies have rapidly generated an unprecedented amount of data that enables us to understand biological systems at single-cell resolution. However, analyzing datasets generated by independent labs remains challenging due to a lack of consistent terminology to describe cell types. Here, we present OnClass, an algorithm and accompanying software for automatically classifying cells into cell types represented by a controlled vocabulary derived from the Cell Ontology. Cell type similarity is inferred according to the distances in the Cell Ontology so a key advantage of OnClass is its ability to annotate cell types that are not present in the training set by using the hierarchical structure of the vocabulary space. We applied OnClass to diverse collections of single cell transcriptomics of both mouse and human and observed substantial improvement on automated cell type annotation. We further demonstrated how OnClass can be used to identify marker genes for cell types present and absent in the training set, suggesting that OnClass can be used as a tool to associate marker genes to each term of the Cell Ontology, offering the possibility of refining the Cell Ontology using a data-centric approach.

Introduction

Single cell RNA-seq has emerged as a powerful tool to generate comprehensive organismal atlases encompassing a wide range of organs and tissues [1–6]. One of the most important tasks in single-cell analysis is cell type annotation, which aims at characterizing and labeling groups of cells according to their gene expression [7–10]. Recent efforts in scRNA-seq have produced an unprecedented large compendium of expert annotated cell types, paving the way for scientists to better understand cellular diversity [4,11]. However, utilizing these cell type annotations is challenging due to the inconsistent terminology used to describe cell types collected by independent groups [4,5,11]. This inconsistency will likely increase as more groups generate new datasets and more cellular types and states are characterized. The Cell Ontology, which used expert curation to organize more than 2000 cell types into a hierarchical structure, offers a controlled vocabulary for cell types and has been proposed as a basis for consistently annotating large-scale single-cell atlases [12–16].

A natural approach to address the inconsistent vocabulary challenge is to build computational methods that automatically assign cells to terms in the Cell Ontology. Ideally, these methods should be fully automated so they can be updated and quickly integrated as the Cell Ontology evolves. However, assigning cells to terms (i.e., cell types) in the Cell Ontology has at least three challenges. First, although the Cell Ontology contains valuable hierarchical relationships among cell types, not all the terms are associated with marker genes which are crucial for cell type annotation. Second, even though supervised learning approaches might be used to predict cell types that have curated annotations, they are unable to classify cells into novel cell types which have been omitted from the training data. Throughout this paper, we refer to “unseen cell types” to describe cell types that are not part of annotated cells in the training data, whereas “seen cell types” are the cell types annotated in the training data. This issue largely prevents us from fully understanding cellular diversity as more than 95% of cell types in the Cell Ontology are unseen even in the largest datasets [11]. Third, as the Cell Ontology is not developed specifically for single-cell RNA sequencing (scRNA-seq), it might be inaccurate for certain cell types relationships. Collectively, these challenges hinder progress towards comprehensive cell type annotation and cellular diversity understanding.

We developed OnClass to address these challenges. OnClass is able to automatically classify cells to any unseen cell type as long as it is in the Cell Ontology. To achieve this, OnClass first infers similarity among all the cell types according to their distances in the Cell Ontology. It then leverages the cell type similarity to transfer annotations to a novel, unseen cell type from other seen and similar cell types. In particular, OnClass constructs a network of cell types based on the hierarchical “is_a” relationship in the Cell Ontology and embeds this network into a low-dimensional space that preserves network topology. It then finds multiple boundaries to separate each cell type into a unique region of the low-dimensional space. Annotated cells and unannotated cells are all mapped to this low-dimensional space. After all cells and cell types are

embedded in the same low-dimensional space, OnClass annotates a cell based on the cell type region in which it lies. Importantly, as this region might correspond to an unseen cell type, OnClass thus enables the annotation of unseen cell types. Furthermore, by embedding cells and the Cell Ontology into the same low-dimensional space, OnClass advances other important applications, such as marker genes identification, data integration, and refinement of the Cell Ontology.

We evaluated OnClass on the Tabula Muris Senis dataset which contains 96 cell types, representing the existing largest effort of cell type characterization. We found that our method outperformed all existing methods and more importantly could classify cells to unseen cell types with more than 0.8 accuracy. We further demonstrated the ability of OnClass to transfer annotations to 26 other single-cell datasets and achieve improved data integration performance, even when the cell types were not part of the training data. Finally, we showed OnClass was able to identify marker genes for both well-characterized cell types as well as unseen cell types, paving the way for creating an organism-wide molecular representation of cellular diversity.

Results

Cell type annotation using OnClass

The Cell Ontology used a controlled vocabulary to organize 2331 cell types into a hierarchy based on the “is_a” relation. OnClass first embedded the Cell Ontology into a low-dimensional space where similar cell types were close to each other [17,18]. It then partitioned this low-dimensional space into multiple regions, each corresponding to a cell type in the Cell Ontology. Single cells, which were characterized by the gene expression, were then projected to this low-dimensional space by finding a nonlinear transformation that projected each cell to the region of its cell type. An unannotated cell can be classified by first projecting to one of the regions in this low-dimensional space using the same nonlinear transformation and then annotating to the corresponding cell type. Importantly, such a classification procedure enables the annotation of unseen cell types based on their regions in the low-dimensional space. In addition to cell type annotation, OnClass used these regions for other applications, including marker genes identification and data integration **(Figure 1a)**.

Cell type embeddings reflect cell type similarity

Since OnClass annotated unseen cell types by transferring annotations from other similar cell types, its performance greatly relied on the quality of the cell type embeddings. High-quality cell type embeddings, which were derived from the Cell Ontology structure, should be similar for cell types with similar gene expression profiles. Therefore, we first verified the merit of our approach by comparing three types of cell type similarities: the Cell Ontology structure-based similarity, the embedding-based similarity, and the gene expression-based similarity (see Online

Methods). We first observed that the embedding-based similarity was strongly correlated with the Cell Ontology structure-based similarity **(Figure 1b)**. For example, the average embedding-based similarity of direct neighbors in the Cell Ontology was 0.86, which was 42% and 183% higher than the average embedding-based similarity of two-hop neighbors and three-hop neighbors. For cell types that are more than four-hop away in the Cell Ontology, the average embedding-based similarity was less than 0.01. Next, we examined whether cell types with similar embeddings would have similar gene expression profiles by comparing the embedding-based similarity and the gene expression-based similarity. Using a collection of annotated cells as the benchmark, we observed strong correlations between these two types of similarities (**Figure 1c,d**). For instance, the correlation between the gene expression-based similarity and the embedding-based similarity was 0.70 (p-value < 1e-10) in pancreas and 0.77 (p-value < 1e-11) in kidney. The strong correlation between these two types of similarities demonstrated the high-quality of cell type embeddings and further suggested the possibility to annotate unseen cell types by transferring annotations from other similar cell types. Unfortunately, none of the existing cell type annotation methods integrates with the Cell Ontology. OnClass's ability to use the Cell Ontology to annotate unseen cell types led us to consider whether we could improve cell type annotation on large and diverse collections of scRNA-seq datasets.

Improved cell type annotation

We ran OnClass on the Tabula Muris Senis (TMS) dataset containing a total of 96 cell types[11]. To investigate the effect of unseen cell types, we split cells into test and training across different proportions of seen cell types in the test set. Overall, we observed that OnClass led to a substantial improvement in comparison to existing approaches (**Figure 2a-d**). We first examined the ability of OnClass to identify cells belonging to a given cell type. We observed that OnClass significantly outperformed all existing approaches in terms of AUROC on all proportions of seen cell types (**Figure 2a**). Even when only half of the cell types were seen, OnClass still achieved an AUROC of 0.87, while AUROCs of existing methods were all below 0.72. Next, we investigated whether OnClass could accurately predict the cell type for a given cell. In a simpler setting where we combined all unseen cell types as an "unseen" class, OnClass outperformed existing methods in terms of Cohen's Kappa statistic (i.e., balanced accuracy) from 10% to 90% unseen cell types (**Figure 2b**). We found that the improvement of OnClass was more prominent with the increasing proportion of unseen cell types. We next evaluated a more challenging setting where unseen cell types were no longer combined and a prediction was deemed as correct only when the cell was assigned to the specific cell type, even if this cell type was an unseen cell type. By using Accuracy@3 and Accuracy@5 to quantify the performance, we observed significant improvement of OnClass in comparison to existing methods (**Figure 2c,d**). For example, when 30% of cell types were unseen, OnClass obtained 0.45 Accuracy@3 and 0.55 Accuracy@5, while none of the existing approaches was greater than 0.3. Again, the improvement of OnClass was larger with more unseen cell types, indicating the advantage of using the Cell Ontology to transfer annotations from seen cell types to unseen cell types.

Notably, even though TMS had one of the most diverse and largest numbers of cell types, it still only covered less than 5% of all cell types in the Cell Ontology. We anticipate that OnClass will be even more useful as more single cell RNA-seq datasets become available.

OnClass accurately annotates unseen cell types

We then examined the performance of OnClass in the more challenging case of annotating unseen cell types, which cannot be achieved by any existing methods. Although recent efforts had classified cells into a "unknown" type [7,8], they could neither break down this new type into detailed subtypes nor annotate it to specific cell types in the Cell Ontology. To enable a better comparison between OnClass and these approaches, we proposed comparison approaches which extended existing approaches by classifying "unknown" type cells to the nearest cell type in the Cell Ontology. We studied the performance of OnClass by using an increasing number of cell types as the training set. We observed significant improvement with OnClass across different proportions of cell types as the training set. For instance, when 60% of cell types were used as training data (**Figure 3a**), OnClass obtained an AUROC of 0.73 when all test cells belonging to unseen cell types, which is 30% higher than comparison approaches. Even when only 20% of cell types are used in the training set, OnClass still obtained an AUROC of 0.68. On a randomly selected set of 9 new unseen types, OnClass was able to accurately classify 81% of cells (**Figure 3b-d**). On a larger set of 21 unseen types, OnClass still accurately classified 58% of cell types (**Figure 3h-j**). We showed the comparison of OnClass annotation and ground truth annotation in **Figure 3b-j**. We found that OnClass was able to accurately classify a majority of cell types, including rare cell types. For those cells that were not accurately annotated, we found that the cell type assigned by OnClass was indeed biologically related to the ground truth cell type. For example, OnClass classified fibroblasts cells as mesenchymal stem cells, which are known to be morphologically indistinguishable from fibroblasts[19] (**Fig. 3h, i**). As human annotation can be imperfect and mostly limited to familiar cell types, OnClass can correct these false positives and broaden expert knowledge.

We next examined the robustness and applicability of OnClass by using it to transfer annotations from TMS to other more diverse datasets across animals, technologies, and organs. In particular, we trained OnClass using TMS and then classified 105,476 cells collected from 26 single-cell datasets (26-datasets) representing 9 technologies and 11 studies. We observed an average AUROC=0.75 for these 26 datasets. Among all 10 cell types, OnClass obtained an AUROC greater than 0.8 for 5 of them (**Figure 4a**). For B cell and macrophage that have annotated cells in TMS, OnClass obtained AUROCs of 0.99 and 0.97, respectively (**Figure 4b, c**). More importantly, for cell types that were unseen in TMS, OnClass still achieved relatively high AUROCs (0.85 for $CD14^+$ monocytes cell, 0.85 for $CD56^+$ natural killer cell, and 0.81 for regulatory T cell), indicating its ability to accurately annotate and discover new cell types (**Figure 4d-f**). Furthermore, the predicted cell type annotations can be used as features to cluster and integrate cells from different datasets. We used the predicted cell type annotations to integrate these 26 datasets following the same procedure as previous work [20]. We observed

good performance by using OnClass, where cells were clustered based on cell types rather than artifacts related to platforms (**Figure 4g**). We further quantified the integration performance using the silhouette coefficient and observed a significant improvement in comparison to the state-of-the-art data integration approach Scanorama [20] (**Figure 4h**), indicating OnClass's robustness to annotating cells from different batches and datasets.

OnClass identifies marker genes for both seen and unseen cell types

Given the accurate annotation of both seen and unseen cell types, we were then interested in using OnClass to identify cell type marker genes. While cell type marker genes are the key to expert curation and understanding cellular diversity, existing knowledge about cell type marker genes is incomplete and limited to extensively studied cell types. Here, we used OnClass to identify marker genes for each cell type, including both seen and unseen cell types in TMS. (**Figure 5a**). For example, OnClass was able to identify marker genes for 64% of seen cell types by examining the top 10 candidate genes in the predicted marker gene list. More importantly, since OnClass did not require any annotations to identify marker genes, it was able to find marker genes for unseen cell types as well. We found that OnClass was able to identify marker genes for 39% of unseen cell types by examining the top 10 candidate genes in the predicted marker gene list. We incorporated these OnClass referred maker genes into the existing Cell Ontology, in the hope of facilitating future expert curation (**Supplementary Table 1**). Although these marker genes were by no means a complete representation of cell type features, they provided a first draft attempt to create a comprehensive characterization of cellular diversity.

Finally, we sought to examine whether our referred marker genes can accurately annotate cells. We first used FACS cells in TMS to identify marker genes and then used them to annotate droplet cells in TMS. We found that the performance of using OnClass referred marker genes was substantially better than using curated marker genes for cell types with more than 500 cells. For example, OnClass marker genes achieved 0.98 AUROC, whereas curated marker genes achieved 0.90 AUROC for cell types with more than 500 and less than 1500 cells (**Figure 5b**). For rare cell types, the performance of OnClass was comparable to curated marker genes (**Figure 5b**). Furthermore, for those cell types that have no curated marker genes, OnClass marker genes also achieved highly accurate cell type annotation performance (**Figure 5c**). For instance, OnClass marker genes obtained a high AUROC of 0.97 for cell types with more than 500 cells. We found that the performance of OnClass was better for cell types with more annotations. As more data would be available in the future, we anticipate further improvement of OnClass on the identification of robust and accurate marker genes. To ensure the robustness of these marker genes, we next used them to classify the 26-datasets. Among all the 10 cell types, 8 of them achieved AUROCs larger than 0.7 and 4 of them achieved AUROCs larger than 0.8 (**Figure 5d-i**). Even for cell types that had no annotation, OnClass still obtained a desirable performance (**Figure 5g, h, i**). Notably, when comparing the performance with a supervised classifier, we found that using marker genes could achieve better results on several cell types (e.g., $CD14^{+}$ monocyte cells) (**Figure 4a, Figure 5a**). Although supervised models are more

expressive, they are also prone to overfitting. In contrast, marker genes are not only interpretable but also more robust to noise, thus enabling accurate annotation of new cells.

Discussion

Ever since the emergence of scRNA-Seq, cell type annotation is a key step in single-cell data analyses. As more cell types are discovered and expected to be discovered, recent efforts have focused on classifying cells into existing labels or a generic unseen cell type [7,8]. Despite encouraging results based on these approaches, these methods fail to provide meaningful information specific to the cell types that are not part of the training sets. In contrast, our method takes an important step forward by mapping each cell to the Cell Ontology, leading to accurate annotations of unseen cell types that cannot be achieved by any existing methods. Conceptually and methodologically, this is substantially different from any existing methods in the sense that our method not only leverages known cell-to-cell (hierarchical) relationships, but also directly classifies cells without pre-clustering in order to model the diversity within unseen cell types.

While our method leverages the Cell Ontology to classify unseen cell types, it is inspired by recent progress in single cell dataset integration approaches [20,21]. In the state-of-the-art single cell integration frameworks, datasets from different technologies are aligned in the same low-dimensional space by using mutual nearest neighbors as anchors to connect them. Indeed, our method can be considered to be aligning the Cell Ontology to the gene expression matrix by using known annotations as anchors. The key novelty of our method comes from effectively embedding cell types based on the Cell Ontology and dividing the low-dimensional space into regions to enable unseen cell type annotation. In future efforts, we might focus on annotating datasets from different technologies simultaneously by mapping all of them into the cell type low-dimensional space, providing a framework that is more applicable, flexible and robust.

With the continually and massively generation of single cell datasets, more cell types will be discovered and annotated. OnClass provides a robust, accurate, efficient and reproducible solution to this problem with a Python-based implementation and an R-based pipeline through the reticulate library. OnClass is publicly available at https://github.com/wangshenguiuc/OnClass under an Open Source software license.

Figure 1. a, Flow chart of OnClass. The Cell Ontology is used to embed cell types into a low-dimensional space. OnClass then finds multiple boundaries in this low-dimensional space to separate each cell type into a unique region. Cells are then projected into this space by reducing the dimensionality of the gene expression matrix. These boundaries can then be used to predict cell type, identify marker genes and integrate datasets. **b**, Violin plot showing the correspondence between the location of each cell type's nearest neighbor in the Cell Ontology and the embedding similarity. The nearest neighbor of each cell type is calculated by using the cosine distance between cell type low-dimensional representations. **c,d** Scatter plots showing the correlations between the embedding-based cell type similarity and the gene expression-based cell type similarity in pancreas (c) and kidney (d).

Figure 2. a-d Bar plots comparing OnClass and existing methods in terms of AUROC (a), Cohen's Kappa (b), Accuracy@3 (c) and Accuracy@5 (d). x-axis shows the proportion of cell types present in the test data.

Figure 3. a, Bar plot comparing OnClass and existing methods for different proportions of seen cells in the training set. x-axis shows the proportion of seen cell types in the training data and y-axis the AUROC. **b,c,e,f,h,i**, 2-D UMAP showing the predicted cell types of OnClass (b, e, h) and ground truth labels (c, f, i) for 9 cell types (b, c), 11 cell types (e, f), and 21 cell types (h, i). The same color means correct annotation. **d,g,j,** Sankey diagrams of the resulting mapping between predicted cell types (left) to ground truth labels (right) for 9 unseen cell types (d), 11 unseen cell types (g) and 21 unseen cell types (j).

Figure 4. a, Bar plot showing the AUROC of OnClass on 9 cell types, including 2 present in TMS (green) and 7 not (yellow). **b-f** AUROC plots of OnClass's prediction for five cell types: B cell (b), macrophage (c), $CD14^+$ monocyte cell (d), $CD56^+$ NK cell (e) and regulatory T cell (f). **g**, 2-D UMAP showing OnClass's integration of 26 datasets on 6 cell types. **h**, Box plot showing the comparison between OnClass and Scanorama on data integration in terms of silhouette coefficient.

Figure 5. a, Plot showing the proportion of cell types out of the ones present (green) or not (yellow) in TMS for which OnClass can identify the marker genes in the top k genes out of 23,437 genes. k is shown in the x-axis and corresponds to the position in the marker gene list sorted by p-value. **b**, Boxplot showing the cell type annotation performance of using OnClass referred marker genes (red) and curated marker genes (blue) in terms of AUROC. x-axis shows the number of cells per cell type. **c**, Boxplot showing the cell type annotation performance of using OnClass referred marker genes in terms of AUROC. Only cell types that have no curated marker genes are shown here. x-axis shows the number of cells per cell type. **d**, Bar plot showing the AUROC of OnClass for 10 cell types, including 2 present in TMS (green) and 8 not (yellow). **e-i** AUROC plots of OnClass's prediction for five cell types: macrophage (e), B cell (f), $CD14^+$ monocyte cell (g), $CD56^+$ NK cell (h), and regulatory T cell (i).

Online methods

scRNA-seq datasets

We used the compendium of single cell transcriptomic data from the Tabular Muris Senis[11]. The cell type annotations in Tabula Muris Senis had been curated by domain experts (**Supplementary Table 2**) and all the cell type annotations present in the dataset were manually mapped to the Cell Ontology vocabulary. We next obtained 26 scRNA-seq datasets from 11 different studies [22–33]. We used the preprocessed collection from Scanorama[20], where low-quality cells were excluded. There were 5,216 genes across all 26 datasets and a total of 105,476 cells, with each dataset containing between 90 and 18,018 cells. Since these datasets did not provide cell type annotations that were mapped to the Cell Ontology vocabulary, we manually mapped cell types in these datasets to the Cell Ontology vocabulary (**Supplementary Table 3**). After the mapping, there were 10 different cell types in these 26 datasets. We denoted these datasets as "26-datasets" in this paper.

The Cell Ontology

We downloaded the Cell Ontology from The OBO Foundry (http://www.obofoundry.org/ontology/cl.html)[13]. We used the "is_a" relation in the Cell Ontology to construct an undirected network of cell types. There were in total of 2331 nodes in the constructed network, corresponding to 2331 different cell types. All edges in this network had the same weight.

Embedding the Cell Ontology into low-dimensional space

OnClass computed a compressed, low-dimensional representation of each cell type based on the constructed cell type network. We used DCA[17,18], which had been proposed to embed the Gene Ontology, to embed the Cell Ontology. DCA first computed a propagated cell type network by applying the random walk with restart[34] to the cell type network. It then obtained the low-dimensional representation of each cell type by using the singular value decomposition (SVD)[35] to reduce the dimensionality of this propagated cell type network. As suggested by DCA, we set the dimensionality of SVD to 1000 and the restart probability of the random walk with restart to 0.8.

Cell type annotation

OnClass used a bilinear model to predict the cell type for a novel cell. Let *M* be an *m* by *n* matrix of input gene expression data, where *m* was the number of cells and *n* was the number of genes. Let *Y* be an m by *c* label matrix, where *c* was the total number of cell types in the Cell Ontology. Y_{ij}=1 if cell *i* belonged to cell type *j*, otherwise Y_{ij}=0. Note that *c* was much larger than the number of seen cell types in the training data, as the majority of cell types were unseen in the training data. For example, there were 96 cell types in TMS, which was much smaller than *c*=2331 cell types in the Cell Ontology. The corresponding columns of unseen cell types were all zeros in the label matrix. Let *U* be a *c* by *q* matrix of the low-dimensional representations of cell types, where *q* was the dimension of cell type low-dimensional space. *U* was the output of DCA and fixed during optimization. OnClass optimized the following cross-entropy loss:

$$L = \Sigma_{i=1}^{m}\Sigma_{j=1}^{c} Y_{ij} log(exp(M_i W_1 W_2 U_j^T) / \Sigma_{k=1}^{c} exp(M_i W_1 W_2 U_k^T)),$$

where $W_1 \in R^{n \times h}$ and $W_2 \in R^{h \times q}$ were the parameters that needed to be estimated. *h* was the hidden-dimension and set to 500. We observed that the performance of OnClass was stable for *h* between 200 and 2000. OnClass used ADAM[36] to optimize this objective function.

After the optimization, the cell type of a new cell with expression vector *z* could then be predicted as:

$$p_j = exp(z W_1 W_2 U_j^T) / \Sigma_{k=1}^{c} exp(z W_1 W_2 U_k^T),$$

where p_j was the probability that this cell belonged to cell type *j*. $P = \{p_1, p_2, ..., p_c\}$ was the probability distribution that this cell belonged to each cell type. Since *c* was the total number of cell types in the Cell Ontology, OnClass could automatically annotate cell types that were not seen in the training data.

Cell type embeddings reflect cell type similarity

We calculated three types of cell type similarities: the Cell Ontology structure-based similarity, the embedding-based similarity, and the gene expression-based similarity. The Cell Ontology structure-based similarity was calculated as the shortest distance between two cell types in the Cell Ontology-based cell type network. The embedding-based similarity was the cosine similarity between low-dimensional representations of two cell types. To calculate the gene expression-based similarity, we used the gene expression of all FACS cells in TMS. The calculation was performed within individual organs. We first identified two sets of cells belonging to two given cell types and an organ. We then calculated the mean of pairwise cosine similarities between these two sets of cells and used it as the gene expression-based cell type similarity.

Evaluation of cell type annotation

We evaluated across different proportions of seen cell types in the test set ranging from 100% to 10%, where 100% indicates that all cell types in the test set also presented in the training set. For a proportion *k*, we first randomly selected *k* percentage of cell types as seen cell types and the remaining as unseen cell types. All cells belonging to these unseen cell types were used as the test set. For the seen cell types, we random split their cells into five equal size folds, where one fold was used as the training set and the remaining four folds were used as the test set. We created a five-fold of test and training here according to the setting of Tabula Muris Senis, where about 20% of cells (3- month mice) were annotated first and then extended to the remaining 80%. The test data thus contained all cells in the unseen cell types and 80% of cells in the seen cell types. We performed cross-validation by repeating this procedure 5 times for each proportion.

To evaluate the case where all cell types in the test set are unseen (**Figure 3a**), we compared the performance across different proportions of seen cell types in the training set. For a given proportion *k*, we randomly selected *k* percentage of cell types as seen cell types and the remaining as unseen cell types. All cells belonging to the seen (unseen) cell types were used as the training (test) set. We performed cross-validation by repeating this procedure 5 times for each proportion.

We evaluated our method and comparison approaches on four metrics, including the area under the receiver operating characteristic curve (AUROC), Accuracy@3, Accuracy@5, and Cohen's kappa statistic[37]. As we were evaluating a large number of classes (i.e., more than 80 cell types), it was important to address the bias from class imbalance during evaluation. Therefore, we used the macro-average AUROC rather than the micro-average AUROC to summarize results across different cell types. Macro-average calculates the areas under the curves for each class independently and then takes the average. Cohen's kappa statistic can handle well both multi-class and imbalanced class problems and has been widely used as an alternative to accuracy. A large cohen's kappa statistic indicates better performance, while 1 indicates perfect classification. Accuracy@3 (Accuracy@5) is a widely used ranking metric, which assesses the correctness of the top 3(5) predicted cell types in comparison to only examining the top 1 cell type in Cohen's kappa statistic. A prediction would be deemed as correct if any of the top 3 (5 for Accuracy@5) predicted cell types is the correct cell type.

Comparison approaches

We compared our method with four existing methods ACTINN, singleCellNet (sCN), one-vs-rest logistic regression (LR), and DOC. ACTINN used a three-layer neural network to predict cell type [9]. We used the implementation of ACTINN from the authors (https://github.com/mafeiyang/ACTINN) and ran it on TMS. We used the default parameters for ACTINN since these parameters were used in their paper to annotate cells in the Tabula Muris[4],

an earlier version of the Tabula Muris Senis. sCN used gene pairs as features and random forest as the classifier to predict the cell type[7]. sCN was able to classify cells into a unknown cell type. We obtained the implementation of singleCellNet from (https://github.com/pcahan1/singleCellNet). We found that the implementation of sCN was not scaled to large datasets like TMS and it was not able to cross-validate rare cell types with less than 50 cells. We reimplemented part of sCN to enable its annotation for rare cell types. To make it scalable to TMS, we ran it on the dimensionality reduced gene expression matrix instead of the original gene expression matrix. LR was the standard machine learning classifier for multi-class classification on large-scale datasets. We used the one-vs-rest logistic regression instead of the multinomial logistic regression in order to obtain a probability cutoff of 0.5 to determine the unknown cell type. DOC was an advanced machine learning method for classifying unseen text documents, which was inherently similar to our problem and could be directly applied here [38]. The key idea of DOC was to find a data-driven probability cutoff for the unknown class rather than using a fixed probability cutoff of 0.5 as LR did. However, DOC was also not able to classify cells into the specific cell type. As the original DOC codebase was developed for word sequences classification and could not take gene expression as input, we reimplemented and replaced its underlying convolutional neural network classifier with a multinomial logistic regression.

Although sCN, DOC and LR were able to classify cells into a “unknown” cell type, they were not able to classify these cells into the specific cell type. To enable a fair comparison, we further proposed to extend these three approaches by classifying each “unknown” cell type cell to a specific cell type. In particular, when a cell was annotated as the “unknown” cell type, we first found the seen cell type that had the largest confidence score for this cell. We then annotated the cell to the nearest neighbor of this seen cell type in the Cell Ontology. We denoted these extended approaches as sCN (extended), LR (extended), and DOC (extended) for sCN, LR, and DOC, respectively.

Transfer annotations to 26-datasets

To transfer annotations from TMS to 26-datasets, we first used Scanorama to correct batch effects among TMS and 26 datasets. Scanorama took the gene expression matrix of these 27 datasets as input, it then provided the corrected gene expression of these 27 datasets. We then ran OnClass on all cells in TMS and predicted cell types for cells in the 26-datasets. To integrate these 26-datasets, we used the output probability distribution of each cell by OnClass as the representation for each cell. We visualized these cells by using UMAP[39] to project these representations. We used silhouette coefficients to evaluate the clustering accuracy for both our method and Scanorama [40].

Marker genes identification

We used differential gene expression analysis to identify marker genes for each cell type. In particular, we first ran OnClass on all FACS cells in TMS and then predicted the probability of these cells belonging to each cell type in the Cell Ontology. For each cell type, we took the 50 cells with the highest probability as the positively annotated group and other 50 cells with the lowest probability as the negatively annotated group. We then used the t-test to test whether an individual gene was significantly overexpressed in the positively annotated group then the negatively annotated group. We performed this one-sided independent t-test for each gene and then ranked genes according to the resulted *P*-values. This rank list was the predicted marker gene list. Curated marker genes of 69 cell types were collected from literature by experts (**Supplementary Table 4**). 28 of 96 cell types in TMS had curated marker genes. To classify a new cell according to marker genes, we used the sum of the expression of marker genes as the predicted score for this cell. A larger score indicated that this cell more likely belonged to the cell type.

Statistical analysis

We used the scipy.stats[41] Python package implementation of the one-sided independent t-test, Pearson correlation statistics, Spearman correlation statistics, and associated p-values used in this study. We used the scikit-learn Python package implementation of one-vs-rest logistic regression, silhouette coefficients, AUROC, and cohen's kappa statistics used in this study[42].

Data availability and code availability

OnClass code and data are available at https://github.com/wangshenguiuc/OnClass

Acknowledgments

The authors would like to thank the developers and maintainers of the Cell Ontology for insightful discussions. This work is supported by the Chan-Zuckerberg Biohub, NIH GM102365, LM005652, and TR002515.

References

1. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
2. Thorsen, T., Roberts, R. W., Arnold, F. H. & Quake, S. R. Dynamic pattern formation in a vesicle-generating microfluidic device. *Phys. Rev. Lett.* **86**, 4163–4166 (2001).
3. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
4. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
5. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **173**, 1307 (2018).
6. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
7. Tan, Y. & Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst* **9**, 207–213.e2 (2019).
8. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
9. Ma, F. & Pellegrini, M. ACTINN: Automated Identification of Cell Types in Single Cell RNA Sequencing. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz592

10. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz292
11. The Tabula Muris consortium *et al.* A Single Cell Transcriptomic Atlas Characterizes Aging Tissues in the Mouse. *bioRxiv* 661728 (2019). doi:10.1101/661728
12. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* **6**, R21 (2005).
13. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
14. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
15. Malladi, V. S. *et al.* Ontology application and use at the ENCODE DCC. *Database* **2015**, (2015).
16. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
17. Wang, S., Cho, H., Zhai, C., Berger, B. & Peng, J. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* **31**, i357–64 (2015).
18. Cho, H., Berger, B. & Peng, J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Syst* **3**, 540–548.e5 (2016).
19. Soundararajan, M. & Kannan, S. Fibroblasts and mesenchymal stem cells: Two sides of the same coin? *J. Cell. Physiol.* **233**, 9099–9109 (2018).

20. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

21. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

22. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

23. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).

24. Grün, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266–277 (2016).

25. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385–394.e3 (2016).

26. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems* **3**, 346–360.e4 (2016).

27. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–31 (2016).

28. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **164**, 325 (2016).

29. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

30. Davie, K. *et al.* A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell* **174**, 982–998.e20 (2018).

31. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).

32. Gierahn, T. M. *et al.* Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).

33. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

34. Pearson, K. The Problem of the Random Walk. *Nature* **72**, 342–342 (1905).

35. Halko, N., Martinsson, P. & Tropp, J. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev.* **53**, 217–288 (2011).

36. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).

37. McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).

38. Shu, L., Xu, H. & Liu, B. DOC: Deep Open Classification of Text Documents. *arXiv [cs.CL]* (2017).

39. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4314

40. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

41. Jones, E., Oliphant, T., Peterson, P. & Others. SciPy: Open source scientific tools for Python. (2001).

42. Kramer, O. Scikit-Learn. in *Machine Learning for Evolution Strategies* (ed. Kramer, O.) 45–53 (Springer International Publishing, 2016).

40. Kramer, O. Scikit-Learn. in *Machine Learning for Evolution Strategies* (ed. Kramer, O.) 45–53 (Springer International Publishing, 2016).