Journal Title Here, 2022, pp. 1–5

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

ArcCell: a robust and generalizable single cell RNA-seq cell type annotation tool

Venkat Suprabath Bitra,[1] Rika Chan[2] and Kristi Xing[2]

[1]School of Engineering and Science, Columbia University, 116th and Broadway, 10027, NY, USA and [2]Biology Department, Barnard College, 3009 Broadway, 10027, NY, USA
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

We introduce ArcCell, a lightweight and generalizable model for single-cell RNA-seq (scRNA-seq) cell type annotation. ArcCell combines balanced pre-processing with ArcFace-based dimensionality reduction and a supervised classifier, achieving strong performance across diverse datasets. On mouse gonadal data, ArcCell reached a macro F1 score of 0.96 on the test set and 0.94 on out-of-sample data, substantially outperforming the CellTypist baseline (F1 scores of 0.62). The model is especially effective for rare cell types, demonstrating high precision and recall. These results highlight ArcCell's robustness and potential for scalable, accurate cell annotation in biological research. The code associated with this work can be found at: https://github.com/VenkatSBitra/coms4761_project

Key words: scRNA-seq, cell type annotation, scRNA-seq preprocessing, dimensionality reduction, rare cell types

Introduction

As single-cell RNA sequencing (scRNA-seq) technology progresses, so does the sheer amount of data produced. The amount of cells reported in publications has been estimated to double each year, the average for 2022 being more than 200,000 cells per publication [9]. As scRNA-seq data continues to grow exponentially, so do the number of cell types, and so arises an increasingly urgent need for automatic cell type annotation.

Existing automatic annotation tools commonly fall into three categories: marker gene database-based, correlation-based, or supervised classification-based [8]. Marker gene database-based models are limited to the discoveries in the current literature, and do not extend to the possibility of conflicts or better existing differential expression in the sequencing data. Seurat is a very commonly used general-purpose scRNA-seq analysis framework, but can generate pseudo-cluster marker genes that are highly enriched in two clusters, due to their one against all approach [11]. Correlation-based models, such as scmap which uses approximate nearest neighbor searches for cluster detection, have issues with scalability and generalizability [6]. scmap doesn't account for batch effects, making it hard to annotate cell types of other datasets. Additionally, especially if doing cell-by-cell comparisons, the scalability is extremely limited. Contrastingly, CellTypist, the inspiration for the creation of this project, is a machine-learning framework that uses logistic regression for automatic cell type classification, and is much more scalable [7]. However, since scRNA-seq is sparse, a highly discriminative dimensionality reduction method may work better for picking out rarer genes, which is the basis of this project.

We introduce ArcCell, a pipeline that includes pre-processing techniques and a machine-learning classifier model that uses a highly discriminative dimensionality reduction method to annotate scRNA-seq data accurately and efficiently. A highly discriminative dimensionality reduction is used due to its improved ability to handle sparse data in combination with pre-processing techniques to balance rare and common cell types (over- and under-sampling, GF-ICF). This model was trained on a male gonadal mouse dataset, and tested using a held-out test sample as well as an out-of-sample set, which was a female gonadal mouse dataset. It was then compared to CellTypist, and manual validation was performed for top contributing genes from ArcCell, CellTypist, and the commonly used Python package scanpy.

Methods

Dataset Description

This study utilizes scRNA-seq datasets derived from mouse gonadal tissue, with a focus on both male and female samples [5]. The datasets were sourced from the cellxgene database, a widely used repository for annotated single-cell data, and exhibit high sparsity with an average of 3,000 detected genes per cell out of 23,333 total measured genes. The primary objective was to enable robust cell type annotation by training and evaluating classification models on these datasets.

The male mouse gonadal dataset was designated as the training set and comprised 30,784 cells, each profiled across 23,333 genes. The female mouse gonadal dataset served as the test set, containing 33,120 cells with the same number of genes.

Both datasets underwent initial quality control to ensure data reliability: cells expressing fewer than 200 genes and genes expressed in fewer than 3 cells were filtered out. Following filtering, the data were normalized and log-transformed to prepare for downstream analysis.

Cell type annotation within these datasets included a diverse range of cell populations. The male dataset encompassed 12 distinct cell types, while the female dataset included 16. Eight cell types were shared between both datasets: endothelial cells, epithelial cells, erythrocytes, germ cells, mesenchymal cells, neural cells, skeletal muscle fiber cells, and supporting cells. This overlap facilitated comparative analysis and model validation across datasets.

The chosen datasets represent developmental gonadal tissue, providing a relevant biological context for studying cell type diversity and annotation methods. The focus on both male and female samples ensures that the resulting models are evaluated for generalizability and robustness across biological variation.

Data Pre-processing

The pre-processing of scRNA-seq data in this project was designed to ensure data quality, enhance the detection of rare cell types, and prepare the data for robust downstream classification. This is especially necessary because there is an inherent class imbalance in scRNA-seq datasets, where common cell types vastly outnumber rare ones. For example, in our male gonadal dataset, there are only 72 erythrocyte cells, compared to 3128 mesenchymal cells, which can lead to overfitting. To mitigate this, both over-sampling and under-sampling strategies were implemented:

1. Under-sampling: Representative data points were identified using KMeans clustering. Around each cluster centroid, random sampling was performed to reduce the number of cells of overrepresented types, aligning their counts with those of rarer cell types.
2. Over-sampling: Synthetic data points were generated for rare cell types by creating weighted averages of neighboring cells in gene expression space. This approach increases the representation of rare cell types without simply duplicating existing data, helping to prevent overfitting.

To further highlight genes important for distinguishing rare cell types, the Gene Frequency-Inverse Cell Frequency (GF-ICF) weighting scheme was applied. The GF-ICF score is computed as follows:

$$\mathrm{GF}_{ij} = \frac{X_{ij}}{\sum_k X_{kj}} \quad \mathrm{ICF}_i = \log\left(\frac{N+1}{n_i+1}\right)$$

$$\mathrm{GFICF}_{ij} = \mathrm{GF}_{ij} \times \mathrm{ICF}_i$$

where X_{ij} is the raw count of gene i in cell j, N is the total number of cells, and n_i is the number of cells in which gene i is detected, i.e., $X_{ij} > 0$. In this method, genes expressed in fewer cells receive higher weights, emphasizing features that may define rare or unique cell populations. This step enhances the model's ability to detect subtle but biologically meaningful differences between cell types.

Dimensionality Reduction

scRNA-seq datasets are inherently high-dimensional, with each cell profiled across thousands of genes. This high dimensionality introduces significant noise and redundancy, which complicates downstream analysis and leads to a phenomenon known as the "curse of dimensionality" [10]. To address this, dimensionality reduction techniques are essential for extracting the most informative features and enabling robust cell type annotation.

In this project, the ArcFace [2] algorithm was employed as the primary dimensionality reduction technique for cell type annotation. Originally developed for face recognition, ArcFace was adapted in this study to operate on gene expression profiles by treating each cell type as a distinct class. ArcFace is a supervised method that projects high-dimensional data onto a lower-dimensional hypersphere while maximizing both intra-class compactness and inter-class separation. By introducing an additive angular margin penalty, ArcFace encourages embeddings of the same cell type to cluster tightly together while pushing different cell types further apart in the embedding space.

The ArcFace model was trained on normalized and log-transformed gene expression data, with class balancing and gene weighting (GF-ICF) applied during pre-processing to enhance the representation of rare cell types. The resulting embeddings were reduced to three dimensions, facilitating both visualization and downstream classification tasks.

While unsupervised methods such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) are commonly used in single-cell analysis for dimensionality reduction and visualization, ArcFace offers a supervised alternative that directly optimizes for cell type discrimination. This approach is particularly advantageous when the goal is accurate cell type annotation, as it leverages known labels during training to improve the separation of biologically meaningful clusters.

Model Architecture and Analysis

The model architecture in this project is centered around a supervised ArcFace-based approach, designed to enhance the discriminative power of cell embeddings in scRNA-seq data. An overview of the architecture is presented in Figure 1. The ArcFace algorithm projects normalized gene expression profiles onto a hypersphere while applying an angular margin penalty, which encourages embeddings from the same cell type to cluster closely together and maximizing separation between different cell types. This is achieved through a softmax output layer that converts these angular margin-based embeddings into cell type probabilities, with cross-entropy loss guiding optimization.

The male mouse gonadal dataset (30,784 cells) was divided into training, validation, and test subsets for model development and evaluation. Data balancing (over- and under-sampling) was performed exclusively on the training split to ensure equal representation of all cell types during model learning. The validation and test splits were reserved for hyperparameter tuning and final performance assessment, respectively. To evaluate generalizability, the female mouse gonadal dataset (33,120 cells) served as an out-of-sample benchmark with partially overlapping cell types. This stratified splitting approach follows best practices in machine learning (ML), preventing data leakage and supporting reliable performance estimation.

For evaluation, the model's performance was assessed using several standard metrics, including accuracy, macro-averaged F1 score, and per-class precision and recall. In addition to overall performance, the model was benchmarked against CellTypist, a logistic regression-based baseline commonly used for automated cell type annotation in single-cell data. This

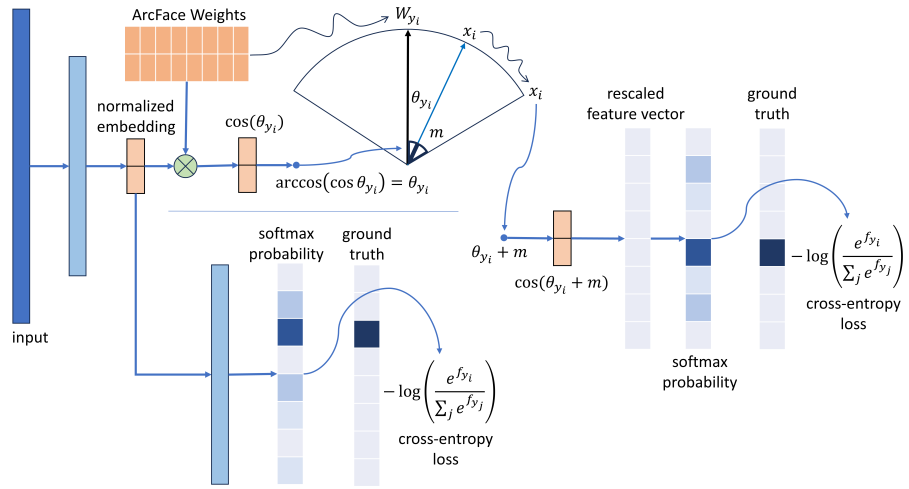

Fig. 1. A schematic of the ArcCell pipeline, illustrating generation of ArcFace embedding and supervised classification steps.

comparative evaluation was conducted using identical training and test splits to ensure fairness.

Interpretability and biological significance of the model learning were addressed by extracting the top contributing genes for each cell type prediction using SHAP (SHapley Additive exPlanations) values, enabling marker gene analysis and biological validation. This comprehensive evaluation framework was designed to address the limitations of correlation-based and traditional ML classifiers, providing a robust and interpretable solution for cell type annotation in complex single-cell datasets.

Results and Discussion

The proposed ArcCell pipeline demonstrated significant improvements in cell type annotation accuracy compared to the CellTypist baseline across multiple evaluation metrics. On the male mouse gonadal test dataset, ArcCell achieved an overall accuracy of 95.73% (F1: 0.9573) versus CellTypist's 58.71% accuracy (F1: 0.6224). This performance gap widened when evaluating generalizability on the out-of-sample female mouse gonadal dataset, where ArcCell maintained 93.51% accuracy (F1: 0.9375) compared to CellTypist's 60.71% (F1: 0.6192).

ArcCell showed particular strength in classifying rare cell types, achieving 94% precision and 94% recall for erythrocytes compared to CellTypist's 12% precision and 99% recall on the same category. For mesenchymal cells–the most abundant type–ArcCell attained 98% precision and 89% recall versus CellTypist's 100% precision but only 21% recall. Additionally, ArcCell achieved superior discrimination across shared cell types, with a macro-averaged F1 score of 0.97 versus CellTypist's 0.66 on test data. Weighted F1 scores (0.96 vs 0.62) confirmed robustness to class imbalance, particularly for rare populations like erythrocytes (94% precision/recall vs 12%/99%).

t-SNE visualization revealed stark differences in embedding quality: ArcCell produced tight, biologically coherent clusters (Figure 2a), while baseline methods showed overlapping cell types (Figure 2b). These visual patterns correlate with quantitative metrics, demonstrating supervised dimensionality reduction enhances both cluster separation and marker gene localization critical for interpretable cell annotation.

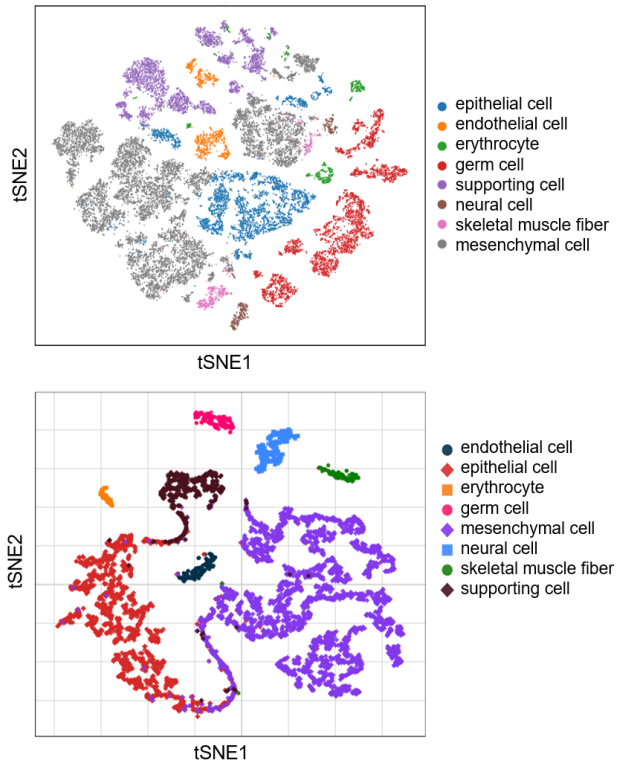

Fig. 2. tSNE plots for the baseline/scanpy (top panel) and ArcCell (bottom panel). Plots are colored by identified cell types on out-of-sample cell types (female gonadal dataset).

Figure 3 summarizes the marker gene analysis comparing ArcCell and the baseline model. Across most cell types, ArcCell successfully identified biologically relevant marker genes. For erythrocytes and neural cells, both models selected *Hbb-a1* and

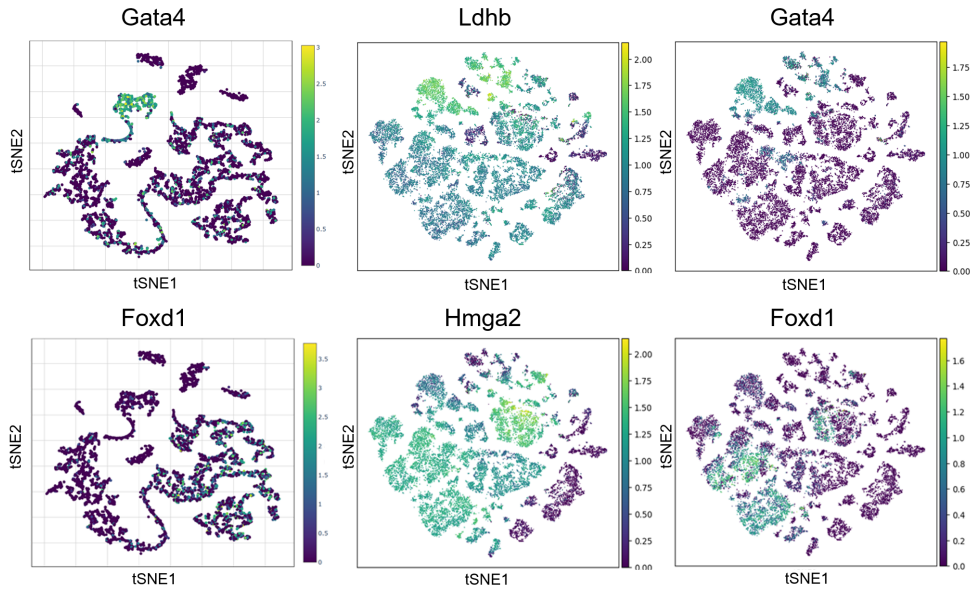

Fig. 3. tSNE plots for top contributing genes identified by the baseline (scanpy) and ArcCell. The first column (Gata4 and Foxd1) are the top genes selected by ArcCell, the second column (Ldhb and Hgma2) are the top genes found by the baseline, and the third column is the differential expression of genes found by ArcCell in the clusters of the baseline. The first row is genes found for supporting cells, and the second row is genes found for mesenchymal cells.

Table 1. Comparative results on test dataset. Evaluation metrics for each cell type in the male gonadal dataset for CellTypist and ArcCell.

		CellTypist			ArcCell		
Cell Type	**Support**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Endothelial cell	167	0.43	1.00	0.60	0.98	1.00	**0.99**
Epithelial cell	1766	0.84	0.69	0.75	0.93	0.94	**0.93**
Erythrocyte	72	0.07	1.00	0.13	1.00	0.99	**0.99**
Germ cell	167	1.00	0.99	**1.00**	1.00	0.99	**1.00**
Mesenchymal cell	3128	1.00	0.37	0.53	0.97	0.96	**0.97**
Neural cell	260	0.80	1.00	0.89	0.99	1.00	**1.00**
Skeletal muscle fiber	124	0.24	0.99	0.39	0.97	0.98	**0.97**
Supporting cell	473	0.41	1.00	0.58	0.93	0.91	**0.92**
Macro Avg	6157	0.60	0.88	0.61	0.97	0.97	**0.97**
Weighted Avg	6157	0.86	0.59	0.62	0.96	0.96	**0.96**
Overall Accuracy			0.5871			**0.9573**	
Overall F1 Score			0.6224			**0.9573**	

Table 2. Comparative results on out-of-sample dataset. Evaluation metrics for each cell type in the female gonadal dataset for CellTypist and ArcCell.

		CellTypist			ArcCell		
Cell Type	**Support**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Endothelial cell	1371	0.73	1.00	0.84	0.98	0.99	**0.99**
Epithelial cell	5322	0.77	0.80	0.79	0.83	0.98	**0.90**
Erythrocyte	911	0.12	0.99	0.21	0.94	0.94	**0.94**
Germ cell	5373	1.00	0.95	0.98	1.00	0.98	**0.99**
Mesenchymal cell	14 074	1.00	0.21	0.34	0.98	0.89	**0.93**
Neural cell	671	0.93	0.99	**0.96**	0.89	1.00	0.94
Skeletal muscle fiber	834	0.18	0.99	0.31	0.57	0.93	**0.71**
Supporting cell	4564	0.86	0.88	0.87	0.98	0.95	**0.97**
Macro Avg	33 120	0.70	0.85	0.66	0.90	0.96	**0.92**
Weighted Avg	33 120	0.89	0.61	0.62	0.95	0.94	**0.94**
Overall Accuracy			0.6071			**0.9351**	
Overall F1 Score			0.6192			**0.9375**	

Ckb as the top markers, respectively, indicating consistency in identifying well-established markers.

ArcCell demonstrated superior marker gene identification for supporting cells by selecting *Gata4* [1], a key regulator of sexual patterning and differentiation into granulosa or Sertoli cells. The baseline model, in contrast, identified *Ldhb*, which is not specific to supporting cells nor directly involved in gonadal development. As shown in Figure 3, ArcCell's embeddings show tight localization of *Gata4* expression within the supporting cell cluster, unlike the baseline, suggesting ArcCell better captures biologically meaningful associations.

For mesenchymal and epithelial cells, ArcCell uniquely identified *Foxd1* [3] as the top marker for both, highlighting its sensitivity to the epithelial-to-mesenchymal transition. The baseline model chose distinct markers, *Hmga2* and *Pax2* [4], which, though relevant, less effectively convey the developmental relationship. ArcCell's tSNE embeddings (Figure 2) distinctly visualize a *Foxd1* expressing cell trajectory and tight clustering, confirming its ability to resolve transitional cell states.

Overall, the marker gene analysis demonstrates that ArcCell not only matches the baseline in identifying canonical markers but also excels at uncovering genes that define rare or transitional cell populations.

Conclusion

In this paper, we introduce ArcCell, a robust and generalizable cell annotation model. Our model is lightweight and scalable, achieving an F1 score of 0.96 compared to CellTypist's 0.62 on the test dataset, and 0.94 vs 0.62 on the out-of-sample dataset. To further validate the accuracy of our results, the top contributing genes to the model were identified using SHAP values. A majority of these genes coincide with marker genes pulled from scanpy, which we use as a baseline for marker gene comparison due to CellTypist's lower annotation accuracy. Additionally, genes that differed from the baseline were primarily rare cell-type markers, which were manually validated as biologically relevant. In conclusion, ArcCell combines pre-processing techniques that balance rare and common cell types in a dataset with a classifier that uses a highly discriminative dimensionality reduction method to annotate cell types effectively and efficiently.

Competing interests

No competing interest is declared.

Author contributions statement

V.B., R.C., and K.X. conceived the experiment. V.B. coded the model, R.C. and K.X. worked on data and pre-processing, R.C. performed manual validation and analyses of marker genes. V.B., R.C., and K.X. wrote the outline presentation, midterm presentation, final presentation, and final report.

Acknowledgments

The authors thank the class, Professor Pe'er, and the TAs for their invaluable feedback on the outline, midterm, and final presentations.

References

1. Malgorzata Bielinska, Amrita Seehra, Jorma Toppari, Markku Heikinheimo, and David B. Wilson. Gata-4 is required for sex steroidogenic cell development in the fetal mouse. *Developmental Dynamics*, 236(1):203–213, 2007.
2. Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, October 2022.
3. Jennifer L. Fetting, Justin A. Guay, Michele J. Karolak, Renato V. Iozzo, Derek C. Adams, David E. Maridas, Aaron C. Brown, and Leif Oxburgh. Foxd1 promotes nephron progenitor differentiation by repressing decorin in the embryonic kidney. *Development*, 141(1):17–27, 01 2014.
4. Juan M. Fons, Oscar H. Ocaña, and M. Angela Nieto. The mutual repression between pax2 and snail factors regulates the epithelial/mesenchymal state during intermediate mesoderm differentiation. *bioRxiv*, 2021.
5. Luz Garcia-Alonso, Valentina Lorenzi, Cecilia Icoresi Mazzeo, João Pedro Alves-Lopes, Kenny Roberts, Carmen Sancho-Serra, Justin Engelbert, Magda Marečková, Wolfram H. Gruhn, Rachel A. Botting, Tong Li, Berta Crespo, Stijn van Dongen, Vladimir Yu Kiselev, Elena Prigmore, Mary Herbert, Ashley Moffett, Alain Chédotal, Omer Ali Bayraktar, Azim Surani, Muzlifah Haniffa, and Roser Vento-Tormo. Single-cell roadmap of human gonadal development. *Nature*, 607(7919):540–547, Jul 2022.
6. Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature Methods*, 15(5):359–362, May 2018.
7. Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.
8. Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scrna-seq data. *Computational and Structural Biotechnology Journal*, 19:961–969, 2021.
9. Laura Tabellini Pierre. Publication trends of single cell rna sequencing research, April 2024.
10. A. Rajkomar, J. Dean, and I. Kohane. Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4:153, 2021.
11. Xin Shao, Jie Liao, Xiaoyan Lu, Rui Xue, Ni Ai, and Xiaohui Fan. Sccatch: Automatic annotation on cell types of clusters from single-cell rna sequencing data. *iScience*, 23(3):100882, Mar 2020.