Integrative and Comparative Biology

Integrative and Comparative Biology, volume 0, pp. 1–9
https://doi.org/10.1093/icb/icae112

Society for Integrative and Comparative Biology

SYMPOSIUM

Exploring the Unknown: How Can We Improve Single-cell RNAseq Cell Type Annotations in Non-model Organisms?

Kevin H. Wong [1], Natalia Andrade Rodriguez and Nikki Traylor-Knowles

Department of Marine Biology and Ecology, Rosenstiel School of Marine, Atmospheric, and Earth Science, University of Miami, Miami, Florida, USA, 33149

From the symposium "Immunity in the 'omics age: what can 'omics approaches tell us about immunity in natural systems?" presented at the annual meeting of the Society for Integrative and Comparative Biology, January 2-6, 2024.

[1]E-mail: kevinhwong1@gmail.com

Synopsis Single-cell RNA sequencing (scRNAseq) is a powerful tool to describe cell types in multicellular organisms across the animal kingdom. In standard scRNAseq analysis pipelines, clusters of cells with similar transcriptional signatures are given cell type labels based on marker genes that infer specialized known characteristics. Since these analyses are designed for model organisms, such as humans and mice, problems arise when attempting to label cell types of distantly related, non-model species that have unique or divergent cell types. Consequently, this leads to limited discovery of novel species-specific cell types and potential mis-annotation of cell types in non-model species while using scRNAseq. To address this problem, we discuss recently published approaches that help annotate scRNAseq clusters for any non-model organism. We first suggest that annotating with an evolutionary context of cell lineages will aid in the discovery of novel cell types and provide a marker-free approach to compare cell types across distantly related species. Secondly, machine learning has greatly improved bioinformatic analyses, so we highlight some open-source programs that use reference-free approaches to annotate cell clusters. Lastly, we propose the use of unannotated genes as potential cell markers for non-model organisms, as many do not have fully annotated genomes and these data are often disregarded. Improving single-cell annotations will aid the discovery of novel cell types and enhance our understanding of non-model organisms at a cellular level. By unifying approaches to annotate cell types in non-model organisms, we can increase the confidence of cell annotation label transfer and the flexibility to discover novel cell types.

Background: Defining cell types through scRNAseq

What is a cell type?

Cells are the basic building blocks of living organisms and the diversity of cells is vast across the animal kingdom with a wide range of functions and phenotypes. Categorizing cells with similar structures and functions into cell types is fundamental to answer key questions in organismal biology, as this shapes our perception and interpretation of how organisms function. Characterizing particular cell groupings reduces the complexity of understanding the cellular compositions of multicellular animals (Arendt 2008), however, the definition of a cell type is subjective and does not follow a unified nomenclature across taxa (Domcke and Shendure 2023). Traditionally, cell types have been described as cells with "hard-wired" characteristics that create specific morphological features corresponding to a particular tissue type (Arendt et al. 2016). However, more recent interpretations of cell types factor in spatial locations within the tissue and molecular aspects that further define cellular function (Wagner, Regev, and Yosef 2016). Consequently, there are continuous discussions on differentiating "cell type" to "cell state" terms due to the complex continuum of cellular characteristics throughout cell developmental trajectories (Trapnell 2015; Domcke and Shendure 2023). Amidst the complexity of defining a cell type, profiling transcriptomes of individual cells, referred to as single-cell RNA sequencing (scRNAseq), has greatly improved our definition of cell types from a molecular perspective across taxa and has enabled discoveries that were pre-

Advance Access publication July 16, 2024

viously limited by bulk RNA sequencing. Further discussions must be continued to delineate novel cell types through scRNAseq in both model and non-model organisms.

Traditional approaches to determine cell types

Cell type discovery can be determined through various approaches that vary in cost and throughput. Cell types are most commonly determined morphologically through approaches such as microscopy (e.g., electron microscopy, high resolution light microscopy) on whole tissue samples (e.g, histology) and cell cultures (Zeng 2022). For example, novel cell types and confirmation of known cell morphologies have been described in the Pacific oyster, *Crassostrea gigas,* through cell cultures and imaging with light microscopy (Potts et al. 2020). However, immortal cell culture lines are not always feasible for all non-model organisms, as difficulties arise in troubleshooting appropriate media and maintaining total viable cell populations for long periods (Roger et al. 2021). Flow cytometry is also a useful tool to describe cell types, as cells can be analyzed by shape, granularity, and size, in addition to specific fluorescence stains that are targets for particular cell characteristics. For example, flow cytometry has been shown to isolate multiple different cell types in human systems (Cossarizza et al. 2021) and novel cell types in non-model marine invertebrates (Snyder et al. 2020). These approaches are foundational to characterize cell types; however, limitations exist, as these methods alone cannot infer functionality, developmental stages, or trajectories of cells within an organism.

Single-cell RNAseq as a tool to determine cell types

Beyond morphological methods, cell types have been classified and discovered through molecular approaches such as scRNAseq and single-nucleus RNA sequencing (snRNAseq). These approaches use transcriptomic profiles of individual cells or nuclei, respectively, to group similar cells by gene expression on a nearest neighbor graph. These high-dimensional cell groupings (i.e., cell clusters) can then be visualized in low-dimensional projections such as Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) or t-distributed Stochastic Neighbor Embedding (t-SNE) (Clarke et al. 2021). Marker genes are then derived by determining genes with significantly higher expression in a specific cluster compared to the other clusters. These analyses are typically performed with programs such as Seurat (Hao et al. 2024), Scanpy (Wolf, Angerer, and Theis 2018), or MetaCell (Baran et al. 2019). From these sets of marker genes, cell cluster annotations can be identified. This is one of the most difficult tasks in scRNAseq data analysis, as it relies on the integration of prior biological knowledge of known cell types with the current data in a reproducible and analytical manner. In model systems, cell type annotation is typically performed through the combination of automatic annotation programs from predefined databases, and manual annotation and validation by experts (reviewed in Clarke et al. 2021). Marker-based annotation methods match known marker genes to a query dataset to manually transfer cell annotation labels. Databases such as Cell Ontology (Diehl et al. 2016), PanglaoDB (Franzén, Gan, and Björkegren 2019), and Cell Marker 2.0 (Hu et al. 2023), are a few resources for manual annotation of cell types by gene markers of model organisms with similar tissue types to humans or mice. Computational programs such as Azimuth (Hao et al. 2021), SingleCellNet (Tan and Cahan 2019), OnClass (Wang et al. 2021), and CellTypist (Xu et al. 2023), can annotate cell types automatically; however, they require a curated database that are typically derived from model organisms. Cell type annotation can also be derived from known gene functions through Gene Ontology (The Gene Ontology Consortium 2018) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa 2008) to understand expressed pathways to hypothesize potential cell types. However, this approach requires a large amount of manual labor, knowledge of specific cell functional characteristics, and complete functionally annotated genomes, which are common limiting factors for non-model organisms.

Cell types have also been further described through multimodal approaches paired with scRNAseq, such as single-cell epigenomics (e.g., Single-cell sequencing assay for transposase-accessible chromatin (scATACseq), single-cell DNA methylation), and spatial transcriptomics (Zeng 2022). Assay for transposase-accessible chromatin using sequencing (ATACseq) is an approach to fragment open regions of chromatin with Tn5 transposases to understand chromatin dynamics and accessible regions of DNA for transcription (Buenrostro et al. 2013). This approach has been adapted for use at a single-cell level, which has enabled researchers to understand how gene regulation through chromatin dynamics can infer cell types (Cusanovich et al. 2018). When paired with scRNAseq, scATACseq is a powerful tool to understand the trajectory of cell types and key regulatory factors that dictate different cell types or states (Ranzoni et al. 2021). Spatial transcriptomics has also been a useful tool to spatially locate cell types in tissues defined by scRNAseq data (Longo et al. 2021). For example, paired scRNAseq and spatial transcriptomic data on mouse brains have allowed a deeper understanding of cell types corresponding to specific brain

regions and also help to determine critical transcription factors that characterize differential brain cell types (Yao et al. 2023). Therefore, the scRNAseq revolution has brought new tools that can enhance our understanding of cell types in combination with other modalities in multicellular organisms.

The power of single-cell sequencing for non-model organisms

Recent developments within scRNAseq have advanced our understanding of the range of cell types present in organisms based on transcriptional diversity. In model systems such as humans and mice, researchers have characterized diverse cell types in multiple organ and tissue types (Tabula Sapiens Consortium∗ et al. 2022). With an ever-increasing availability of diverse scRNAseq datasets, comparisons can now be made across species to understand the evolution of specific cell types. For example, seven different vertebrate species single-cell atlases were compared to understand the evolution of immune cells and molecules across taxa (Jiao et al. 2024). This comparative analysis across high- to low-level vertebrates identified conserved and unique gene markers of both adaptive and innate immune cells, allowing for a fundamental insight into the variability of immune cells across species levels (Jiao et al. 2024). Within non-model invertebrate species, scRNAseq is quickly developing into a powerful tool for identifying novel cell types and characterization of gene expression of specific cell types. For example, within the Caridean shrimp, *Marsupenaeus japonicus,* single-cell expression of hemocytes has allowed the discovery of six subpopulations of hemocytes, increasing the resolution of the previously determined two subpopulations defined morphologically by flow cytometry (Koiwai et al. 2021). In the soft coral, *Xenia sp.*, authors traced the developmental lineage of algal hosting cells through scRNAseq, providing key insight into the mechanisms for coral endosymbiosis uptake and loss (Hu et al. 2020). Transcriptomes of single nuclei isolated from Atlantic salmon (*Salmo salar* L.) livers infected with a common bacterial pathogen *Aeromonas salomonicida*, has allowed us to understand how metabolic remodeling of immune and non-immune cells occurs during pathogenic infection of this important fisheries species (Taylor et al. 2022). Additionally, the authors note that the mammalian hepatic cell marker genes did not translate well to *S. salar* cell markers, therefore outlining the need for better cell marker annotation resources for non-model organisms (Taylor et al. 2022). Single-cell RNA sequencing is a promising and powerful tool for biologists that will complement traditional cell biology approaches; however, careful consideration must be taken while analyzing and interpreting this data type for non-model organisms.

Current limitations in single-cell sequencing for determining cell types in non-model organisms

As scRNAseq is a quickly developing technology that was initially designed for model organisms, we must understand the limitations of this technology while applying it to non-model organisms to be confident in our interpretations. A thorough review by Alfieri et al. (2022) highlights the laboratory challenges of scRNAseq cell capture and library preparation on cells derived from non-model organisms, as they present significant challenges that model organisms (i.e., mammalian) cells do not face. On the computational side, even more challenges are presented due to the lack of genomic resources and knowledge of cell types beyond mammals. High-quality genomic references are not always available for non-model organisms, leading to poor alignment or annotations during data analysis and limiting the resolution of any molecular analysis (Cleves et al. 2020). Additionally, there is not always a straightforward way to translate cell markers from model organisms to non-model organisms especially when one-to-one orthologs do not exist between distantly related species (Nehrt et al. 2011). Ortholog matching to determine cell markers between model and non-model organisms is possible with programs such as OrthoFinder (Emms and Kelly 2019); however, these have limitations and must be used with caution. Orthology inference can be challenging as distantly related species have few one-to-one ortholog comparisons due to variations in gene lengths, rapidly evolving orthogs (Steenwyk et al. 2023), and incomplete genome annotations. Therefore, alternative approaches must be additionally used to validate cell types in non-model organisms, which will enable the discovery of novel species-specific cell types and fully utilize the power of scRNAseq on non-model organisms.

In the remainder of this perspective, we will discuss new methods to increase the confidence of cell type annotations in non-model organisms through scRNAseq. We highlight the potential ways to enhance the discovery of new cell types for non-model organisms by discussing approaches that do not solely rely on genomic annotation resources (Fig. 1). We will discuss topics relating to (i) sequence and lineage-based clustering of cell types across species, (ii) the use of deep learning and artificial intelligence (AI) for cell type classification, and (iii) dark genes as candidate cell markers for non-model organisms.

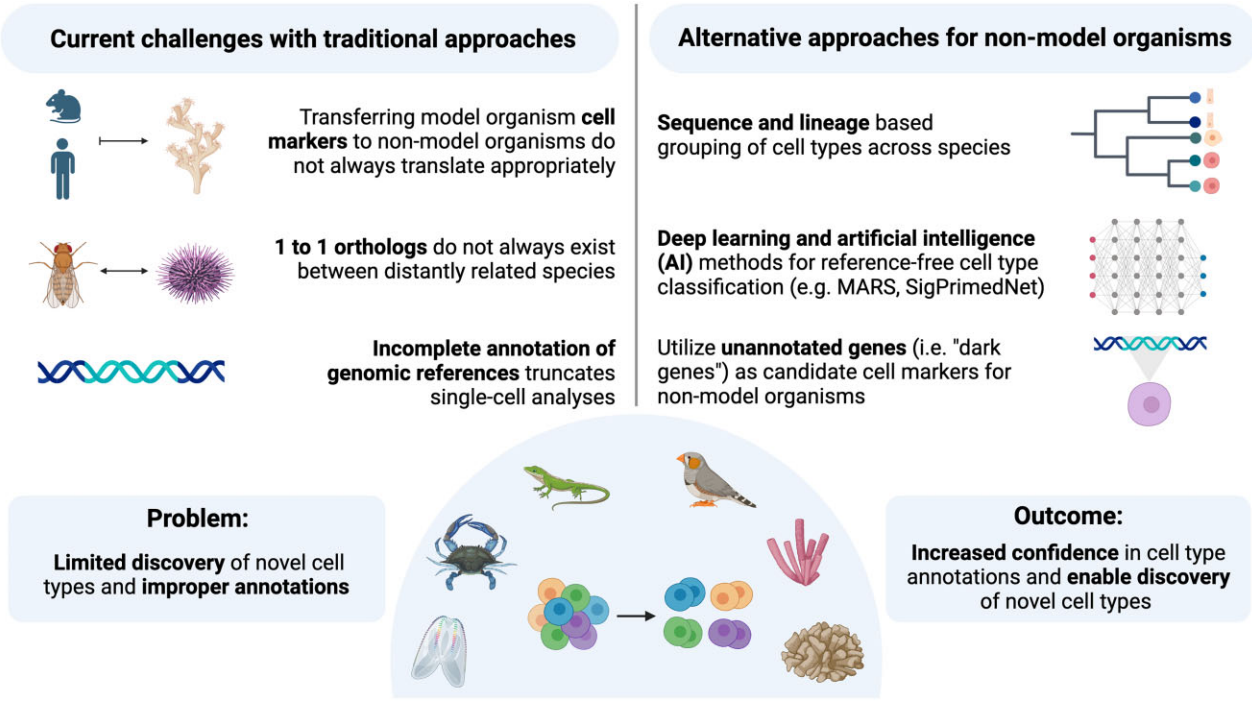

Fig. 1 Summarizing the current challenges and alternative approaches for annotating single-cell RNAseq derived cell types in non-model organisms. This figure was made in BioRender.com.

Alternative approaches to type cell clusters

Cell typing through species comparisons

With the increase of available scRNAseq datasets, integration across species has become a useful approach to not only identify cell types, but also to understand the evolution of species-specific cell types (Song et al. 2023). Mapping single-cell transcriptomes across a variety of taxa is inherently challenging, as (i) many distantly related species do not have one-to-one orthologs due to frequent gene losses and acquisitions over evolutionary history (Nehrt et al. 2011), (ii) marker gene expression similarity is low between taxa, as gene regulation most likely varies from species to species, and (iii) technical batch effects due to sequencing of different single-cell platforms may cause biases between while cell types are captured (Tarashansky et al. 2021). However, new computational methods are being developed to resolve these issues and increase the confidence of comparing single-cell atlases between species.

Programs such as SAMap (Self-Assembling Manifold mapping) can align similar cell types across species and tissue types from scRNAseq datasets (Tarashansky et al. 2021). SAMap can overcome many of these challenges by accounting for the complexities of gene evolution by aligning atlases at both a gene (i.e., marker) and cell (i.e., transcriptome) level. In brief, SAMap uses ortholog information to select conserved genes between datasets that will serve as anchors for downstream dataset alignment. Once the high-dimensional gene expression data are reduced to a lower-dimensional space, mutual nearest neighbors (MNNs) between cells in different datasets can be identified. MNNs are pairs of cells from different datasets that are each other's nearest neighbors in the reduced space, helping to establish correspondences between datasets. Using the MNNs, SAMap then constructs a joint representation of the cells from the different datasets, resulting in a combined manifold that aligns similar cell types across species. Through these methods, SAMap enables various analyses, including the identification of conserved cell type families across species, examination of paralog substitution events, and the study of single-cell evolutionary processes (Tarashansky et al. 2021).

SAMap has been applied on multiple non-model species to further annotate cell atlases. For example, SAMap has helped to annotate novel cell clusters of the freshwater sponge *Spongilla lacustris*, which provided information to distinguish secretory neuroid cells, indicating a sophisticated communication system organized around the sponge's digestive chambers (Musser et al. 2021). These findings suggest that the communication system in sponges uses conserved gene sets that became part of the pre- and post-synapse in the nervous systems of more complex animals, supporting the hypothesis that sensory cells and myocytes may have

evolved from shared ancestral cell types in early metazoans (Musser et al. 2021). In axolotls (*Ambystoma mexicanum*), SAMap was used to compare single cell transcriptomes of five vertebrate species to determine the presence of apical-ectodermal-ridge cells that are crucial for limb development and regeneration. This cross-species comparison determined that *A. mexicanum* do not have apical-ectodermal-ridge cells, but do have cells that perform similar roles in limb regeneration, finally settling the debate of apical-ectodermal-ridge cells in this species (Zhong et al. 2023). Tools such as SAMap will be powerful to annotate cell types in non-model species and to understand cell type diversity across taxa. SAMap is publicly available to use here: https://github.com/atarashansky/SAMap.

An alternative method to annotating cell clusters is through a marker-free phylogenetic approach that classifies cells by topology. Topology elucidates the evolutionary history of cell types while comparing multispecies single-cell datasets, where interior nodes depict ancestral cell types, and exterior node tips depict species-specific cell types (Mah and Dunn 2024). The authors indicate that the principal components of single-cell gene expression data can classify or group cell type clades without the use of cell marker genes, thus resolving the issue of using model organism gene markers on non-model organisms (Mah and Dunn 2024). To briefly understand this approach, scRNAseq datasets are integrated, then a Principal Component Analysis (PCA) is performed on the integrated matrix to reduce dimensionality and identify the most significant features. Phylogenetic trees are then constructed for each PC range using the continuous maximum likelihood method (contml) and evaluated using bootstrap analysis to determine robustness. To identify the best tree a jumble analysis is performed that reflects the robustness of the nodes. Then "average" cells can be created by averaging the PCA values for each cell type. These averaged cells are used to construct a new tree to provide a more generalized view of cell type evolution (Mah and Dunn 2024). A similar concept has been applied to the scRNAseq cell atlas of the mouse nervous system to categorize neural and neuronal cells, then further subclasses of neuron cells (Zeng 2022). This transcriptomic cell type taxonomy allows the authors to determine key genes that are responsible for each taxonomic branch, enabling further insight into the main cell marker genes for each subclass of neurons (Zeng 2022). Therefore, utilizing a marker-free phylogenetic approach may be an alternative method to classify cell types, determine novel cell markers from these cell clade classifications, and understand the evolutionary lineages of cell types in non-model organisms. Code to replicate this analysis can be found here: https://github.com/dunnlab/cellphylo.

These cross-species approaches are just two examples of how we can transfer cell type labels conservatively, which still allows for discovering new cell types and does not necessarily require complete functionally annotated genomes. Approaches that incorporate the evolutionary history of cell types will be critical in classifying cell types within non-model organisms and provide further insight on the discovery of novel and species-specific cell types.

Cell typing with machine learning

Machine learning methods are increasingly used to analyze highly complex genomic data (Wang et al. 2023). Deep learning is a type of machine learning that uses multilayer deep neural networks (DNN) and allows the modeling of highly complex data, such as cell clustering, to improve predictions of clusters or associations. These predictions are possible due to the ability of the DNN to learn patterns of special associations from training data (Eraslan et al. 2019). Multiple deep learning approaches have been tested on single-cell transcriptomic data, showing promising results for classifying cell types. However, overfitting the noisy nature of scRNAseq data remains a problem (Le et al. 2022). Many machine learning tools for single-cell data use DNN models with annotated training sets to classify and annotate cell clusters (Premkumar et al. 2024). While these approaches can be useful for model systems, this approach limits the discovery of novel cell types by excluding cell clusters that are not present in the training dataset. We will discuss the most recent approaches that have been developed to tackle this problem and could be applied to non-model systems.

Meta-learning is a deep learning concept that uses neural networks to learn patterns from training data and then apply that learning while classifying new data with similar characteristics (Hospedales et al. 2022). This method can be applied to scRNAseq data to train the algorithm with a mix of well-annotated model organism cell types and partially annotated non-model organism cell types. This concept increases the confidence of the annotation of non-model organism cell types and allows for new cell type discovery. MARS, is one example of a tool using a meta-learning machine learning approach that can be used to annotate known and unknown cell types in heterogeneous scRNAseq datasets (Brbić et al. 2020). MARS utilizes annotated and unannotated scRNAseq transcriptomes as inputs for meta-data created by using DNN. This approach allows the identification of similar cell types to be em-

bedded close to each other, while different cell types are further apart. MARS has been used to identify concurrent and novel cell types across tissue types and over time series experiments (Brbić et al. 2020). For example, MARS has been implemented on *Drosophila* single-cell transcriptomes to understand the role of various neuron types throughout development. Through the MARS approach, signals of specific neuronal expression during development were identified to specific adult-stage sensory responses (McLaughlin et al. 2021). MARS is an excellent machine learning tool to integrate multispecies datasets to transfer cell type labels without excluding novel cell types. Code to run MARS can be found here: https://github.com/snap-stanford/mars.

Another example of a machine learning approach is SigPrimedNet, an artificial neural network that can annotate known cell types in addition to identifying unknown cell types (Gundogdu et al. 2023). SigPrimedNet integrates domain-specific insights based on KEGG identifiers to the neural network to overcome the constraints associated with traditional supervised clustering methods. Additionally, SigPrimedNet produces low-false positive rates on unknown cell-type annotations through an anomaly detection method. This is the first supervised approach that uses domain-informed gene-to-gene interaction based on KEGG pathways through sparse neural networks to determine unknown cell types (Gundogdu et al. 2023). One drawback of this program is that the network building and training is based on KEGG annotations that are often missing for non-model organism data sets, limiting the use of this tool to only well annotated genomes. However, SigPrimedNet is still a powerful tool for non-model organisms with annotated genomes, as it includes unknown cell types into cell identification approaches. SigPrimedNet can be run with the following open access code here: https://github.com/babelomics/sigprimednet.

Identifying cell types is still a large problem in the single-cell community, as highly dimensional and variable data can be difficult to tease apart in order to form meaningful biological interpretations. Here, we provided a few of the most recent machine learning approaches that can accelerate the cell type discovery process. These technologies alone cannot confidently declare a new cell type and traditional methods must be used to validate. However, machine learning models could help guide research focus on potentially interesting and novel cell types as shown from the examples earlier. As the community moves forward, advanced computational approaches like deep learning may provide unbiased ways of discovering novel cell types in non-model organisms.

Dark genes as cell markers

The concept of "dark genes" in non-model organisms has risen due to the consistent lack of functional annotation in transcriptomic studies. Dark genes can be defined as genes with no functional annotation yet are differentially expressed under specific conditions of an organism, leaving large gaps while interpreting gene expression results (Cleves et al. 2020). Although dark genes have been described at a bulk transcriptomic level, a similar concept can be applied to scRNAseq data. Typically, only genes with known functions are reported as cell markers due to the associated functionality with that specific cell type. Additionally, previous studies cross-annotate cell types based on similar gene functions across datasets, thus excluding potential strong cell markers that are dark genes. Discarding dark genes as cell markers limits the interpretability of cell type annotations and the discovery of novel species-specific cell types. For example, in the sponge *Amphimedon queenslandica*, only three major cell types were identified through scRNAseq based on marker gene annotations and relating to known function, thus limiting the discovery of cell subtypes or other cell types (Sebé-Pedrós et al. 2018). In the ctenophore *Mnemiopsis leidyi,* many of the cell clusters were not annotated due to the lack of annotation on most marker genes. Even with the clusters that were annotated, the authors had low-confidence as some cell markers did not correspond to known tissue types in ctenophores. For example, one of the cluster marker genes was striated-type myosin II, which is a smooth muscle cell marker. However, *M. leidyi* lacks this type of tissue, therefore transferring annotation of cell markers from distantly related species could potentially lead to the misidentification of cell types (Sebé-Pedrós et al. 2018). This outlines the importance of finding alternative methods to infer cell types through marker genes with no annotations to avoid erroneous annotations of cell clusters and characterize novel cell types.

As annotating genes with no known function is a challenging task, we suggest two main approaches for validating dark genes as cell markers (i) flow cytometry to sort potentially novel cell types and (ii) *in situ* hybridization or spatial transcriptomics to visually locate the cells in the animal tissue to infer potential function. Flow cytometry can be used to sort cells based on size, shape, and fluorescence to phenotype cell types before single-cell sequencing, or for validation. For example, breast cancer tumor cells were tagged with monoclonal antibodies and sorted with fluorescence-activated cell sorting (FACS) before bulk RNA sequencing to understand transcriptional profiles (Porter et al. 2020). In non-model systems, similar techniques can be applied

to dark genes associated with novel cell types. We could develop fluorescent probes to target dark gene markers (i.e., fluorescence *in situ* hybridization RNA probes) and sort-tagged populations through FACS. This sorted cell population could then be imaged to confirm the efficacy and confidence of the specificity of the dark gene cell marker for a unique cell type. This has been proven to be useful in isolating human enteric nervous system cells by creating a custom gene probe panel and sorting through FACS (Windster et al. 2023). Therefore, cell population specific sorting through FACS to validate scRNAseq dark gene cell markers will be a powerful tool and prove novel insights into cellular subtypes or developmental stages.

In most single-cell studies, fluorescence RNA *in situ* hybridization (FISH) is a standard in visualizing described genes. In the soft coral *Xenia sp.*, algal hosting cells were visualized with FISH probing for *LePin*, a gene suggested to be selective for Symbiodiniaceae, the common endosymbiont of cnidarians (Hu et al. 2020). FISH could be an essential approach to characterize cell types in non-model systems by using dark genes as probes for specific cell types derived from scRNAseq analyses. This method allows spatial localization and provides tissue-specific context to a specific set of cells that might help elucidate their cell type annotations. The integration of scRNAseq with spatial transcriptomics has been recently used to understand the distribution of cell types and cell-cell communication, thus providing important biological context to the cells defined by scRNAseq clusters (Longo et al. 2021). Therefore, spatially resolved scRNAseq cell type clusters could be further annotated by the visualization of dark gene markers in tandem with spatial transcriptomics.

Conclusion and future directions

As the use of scRNAseq on non-model systems continues to enhance our understanding of cells, it is important to have a consensus with our approaches to rigorously annotate cell types without the exclusion of novel cell type discovery. In this perspective, we provided a few alternative approaches to aid researchers in cell type annotations. By utilizing reference-free approaches, we can fully utilize scRNAseq data by describing cell types not necessarily associated with model systems. As this is a quickly developing field, we expect that there will be more approaches available that can be included in further discussion. Moving forward, we propose further questions that will continue this discussion and where we anticipate further research to be directed:

(1) Can we create a database of gene and ortholog cell markers for non-model organisms?

(2) How can we consolidate all the bioinformatic pipelines used for cell type annotations?

(3) Can we develop a rigorous pipeline to accurately annotate cell types through scRNAseq and other approaches with a community consensus?

With these considerations and questions in mind, we hope to provide the non-model cell biology community with resources and ideas to enhance the scRNAseq cell cluster annotation process to increase cell type discovery and confidence in annotation labels.

Author contributions

K.H.W., N.A.R.: Conceptualization; K.H.W.: Manuscript writing; and K.H.W., N.A.R., and N.T.-K.: Manuscript editing.

Acknowledgments

We thank Dr. Lauren Fuess for organizing the "Immunity in the "omics age: what can" omics approaches tell us about immunity in natural systems?" symposia. We would also like to thank Danielle Becker and the anonymous reviewers for their helpful comments on this manuscript. Figures were created in BioRender.com..

Funding

This work is funded by the National Science Foundation (NSF) Enabling Discovery through GEnomics (EDGE) grant [2128071] awarded to N.T.-K.

Conflict of interest

All authors delcare no conflicts of interest.

References

Alfieri JM, Wang G, Jonika MM, Gill CA, Blackmon H, Athrey GN. 2022. A Primer for Single-Cell Sequencing in Non-Model Organisms. Genes 13:380. https://doi.org/10.3390/genes13020380.

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, et al. 2016. The Origin and Evolution of Cell Types. Nat Rev Genet 17:744–57. https://doi.org/10.1038/nrg.2016.127.

Arendt D. 2008. The Evolution of Cell Types in Animals: emerging Principles from Molecular Studies. Nat Rev Genet 9:868–82. https://doi.org/10.1038/nrg2416.

Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, Meir Z, Hoichman M, Lifshitz A, Tanay A. 2019. MetaCell: analysis of Single-Cell RNA-Seq Data Using K-Nn Graph Partitions. Genome Biol 20:206. https://doi.org/10.1186/s13059-019-1812-2.

Brbić M, Zitnik M, Wang S, Pisco AO, Altman RB, Darmanis S, Leskovec J. 2020. MARS: discovering Novel Cell Types across Heterogeneous Single-Cell Experiments. Nat Methods 17:1200–6. https://doi.org/10.1038/s41592-020-00979-3.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. Nat Methods 10:1213–8. https://doi.org/10.1038/nmeth.2688.

Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, Macparland SA, Bader GD. 2021. Tutorial: guidelines for Annotating Single-Cell Transcriptomic Maps Using Automated and Manual Methods. Nat Protoc 16:2749–64. https://doi.org/10.1038/s41596-021-00534-0.

Cleves PA, Shumaker A, Lee J, Putnam HM, Bhattacharya D. 2020. Unknown to Known: advancing Knowledge of Coral Gene Function. Trends Genet 36:93–104. https://doi.org/10.1016/j.tig.2019.11.001.

Cossarizza A, Chang H-D, Radbruch A, Abrignani S, Addo R, Akdis M, Andrä I, Andreata F, Annunziato F, Arranz E, et al. 2021. Guidelines for the Use of Flow Cytometry and Cell Sorting in Immunological Studies (third Edition). Eur J Immunol 51:2708–3145. https://doi.org/10.1002/eji.202170126.

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, Dewitt WS, et al. 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174:1309–24.e18. https://doi.org/10.1016/j.cell.2018.06.052.

Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, He Y, Osumi-Sutherland D, Ruttenberg A, Sarntivijai S, et al. 2016. The Cell Ontology 2016: enhanced Content, Modularization, and Ontology Interoperability. J Biomed Semant 7:44. https://doi.org/10.1186/s13326-016-0088-7.

Domcke S, Shendure J. 2023. A Reference Cell Tree Will Serve Science Better than a Reference Cell Atlas. Cell 186:1103–14. https://doi.org/10.1016/j.cell.2023.02.016.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic Orthology Inference for Comparative Genomics. Genome Biol 20:1–14. https://doi.org/10.1186/s13059-019-1832-y.

Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep Learning: new Computational Modelling Techniques for Genomics. Nat Rev Genet 20:389–403. https://doi.org/10.1038/s41576-019-0122-6.

Franzén O, Gan L-M, Björkegren JLM. 2019. PanglaoDB: a Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data. Database: The Journal of Biological Databases and Curation 2019:1–9. https://doi.org/10.1093/database/baz046.

Gundogdu P, Alamo I, Nepomuceno-Chamorro IA, Dopazo J, Loucera C. 2023. SigPrimedNet: a Signaling-Informed Neural Network for scRNA-Seq Annotation of Known and Unknown Cell Types. Biology 12:579. https://doi.org/10.3390/biology12040579.

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al. 2021. Integrated Analysis of Multimodal Single-Cell Data. Cell 184:3573–87.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, et al. 2024. Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis. Nat Biotechnol 42:293–304. https://doi.org/10.1038/s41587-023-01767-y.

Hospedales T, Antoniou A, Micaelli P, Storkey A. 2022. Meta-Learning in Neural Networks: a Survey. IEEE Trans Pattern Anal Mach Intell 44:5149–69.

Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, Chen J, Jiang W, Yang K, Ou Q, et al. 2023. CellMarker 2.0: an Updated Database of Manually Curated Cell Markers in Human/mouse and Web Tools Based on scRNA-Seq Data. Nucleic Acids Res 51:D870–6. https://doi.org/10.1093/nar/gkac947.

Hu M, Zheng X, Fan C-M, Zheng Y. 2020. Lineage Dynamics of the Endosymbiotic Cell Type in the Soft Coral Xenia. Nature 582:534–8. https://doi.org/10.1038/s41586-020-2385-7.

Jiao A, Zhang C, Wang X, Sun L, Liu H, Su Y, Lei L, Li W, Ding R, Ding C, et al. 2024. Single-Cell Sequencing Reveals the Evolution of Immune Molecules across Multiple Vertebrate Species. J Adv Res 55:73–87. https://doi.org/10.1016/j.jare.2023.02.017.

Kanehisa M. 2008. The KEGG Database. "In Silico" Simulation of Biological Processes. West Sussex: John Wiley & Sons, Ltd. p. 91–103.

Koiwai K, Koyama T, Tsuda S, Toyoda A, Kikuchi K, Suzuki H, Kawano R. 2021. Single-Cell RNA-Seq Analysis Reveals Penaeid Shrimp Hemocyte Subpopulations and Cell Differentiation Process. eLife 10:e66954. https://doi.org/10.7554/eLife.66954.

Le H, Peng B, Uy J, Carrillo D, Zhang Y, Aevermann BD, Scheuermann RH. 2022. Machine Learning for Cell Type Classification from Single Nucleus RNA Sequencing Data. PLoS One 17:e0275070. https://doi.org/10.1371/journal.pone.0275070.

Longo SK, Guo MG, Ji AL, Khavari PA. 2021. Integrating Single-Cell and Spatial Transcriptomics to Elucidate Intercellular Tissue Dynamics. Nat Rev Genet 22:627–44. https://doi.org/10.1038/s41576-021-00370-8.

Mah JL, Dunn CW. 2024. Cell Type Evolution Reconstruction across Species through Cell Phylogenies of Single-Cell RNA Sequencing Data. Nat Ecol Evol 8:325–38.

Mclaughlin CN, Brbić M, Xie Q, Li T, Horns F, Kolluru SS, Kebschull JM, Vacek D, Xie A, Li J, et al. 2021. Single-Cell Transcriptomes of Developing and Adult Olfactory Receptor Neurons in. eLife 10:e63856. https://doi.org/10.7554/eLife.63856.

Musser JM, Schippers KJ, Nickel M, Mizzon G, Kohn AB, Pape C, Ronchi P, Papadopoulos N, Tarashansky AJ, Hammel JU, et al. 2021. Profiling Cellular Diversity in Sponges Informs Animal Cell Type and Nervous System Evolution. Science 374:717–23. https://doi.org/10.1126/science.abj2949.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. PLoS Comput Biol 7:e1002073. https://doi.org/10.1371/journal.pcbi.1002073.

Porter W, Snowden E, Hahn F, Ferguson M, Tong F, Dillmore WS, Blaesius R. 2020. High Accuracy Gene Expression Profiling of Sorted Cell Subpopulations from Breast Cancer PDX Model Tissue. PLoS One 15:e0238594. https://doi.org/10.1371/journal.pone.0238594.

Potts RWA, Gutierrez AP, Cortés-Araya Y, Houston RD, Bean TP. 2020. Developments in Marine Invertebrate Primary Culture Reveal Novel Cell Morphologies in the Model Bivalve. PeerJ 8:e9180. https://doi.org/10.7717/peerj.9180.

Premkumar R, Srinivasan A, Harini Devi KG, M D, E G, Jadhav P, Futane A, Narayanamurthy V. 2024. Single-Cell Classification, Analysis, and Its Application Using Deep Learning Techniques. Biosystems 237:105142. https://doi.org/10.1016/j.biosystems.2024.105142.

Ranzoni AM, Tangherloni A, Berest I, Riva SG, Myers B, Strzelecka PM, Xu J, Panada E, Mohorianu I, Zaugg JB, et al. 2021. Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell 28:472–87.e7. https://doi.org/10.1016/j.stem.2020.11.015.

Roger LM, Reich HG, Lawrence E, Li S, Vizgaudis W, Brenner N, Kumar L, Seetharaman JK, Yang J, Putnam HM, et al. 2021. Applying Model Approaches in Non-Model Systems: a Review and Case Study on Coral Cell Culture. PLoS One 16:e0248953.

Sebé-Pedrós A, Chomsky E, Pang K, Lara-Astiaso D, Gaiti F, Mukamel Z, Amit I, Hejnol A, Degnan BM, Tanay A. 2018. Early Metazoan Cell Type Diversity and the Evolution of Multicellular Gene Regulation. Nat Ecol Evol 2:1176–88.

Snyder GA, Browne WE, Traylor-Knowles N, Rosental B. 2020. Fluorescence-Activated Cell Sorting for the Isolation of Scleractinian Cell Populations. J Vis Exp 159: e60446. https://doi.org/10.3791/60446.

Song Y, Miao Z, Brazma A, Papatheodorou I. 2023. Benchmarking Strategies for Cross-Species Integration of Single-Cell RNA Sequencing Data. Nat Commun 14:6495. https://doi.org/10.1038/s41467-023-41855-w.

Steenwyk JL, Li Y, Zhou X, Shen X-X, Rokas A. 2023. Incongruence in the Phylogenomics Era. Nat Rev Genet 24:834–50. https://doi.org/10.1038/s41576-023-00620-x.

Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaup B, Brown P, et al. 2022. The Tabula Sapiens: a Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans. Science 376:eabl4896.

Tan Y, Cahan P. 2019. SingleCellNet: a Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Syst 9:207–13.e2. https://doi.org/10.1016/j.cels.2019.06.004.

Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, Wang B. 2021. Mapping Single-Cell Atlases throughout Metazoa Unravels Cell Type Evolution. eLife 10:e66747. https://doi.org/10.7554/eLife.66747.

Taylor RS, Ruiz Daniels R, Dobie R, Naseer S, Clark TC, Henderson NC, Boudinot P, Martin SAM, Macqueen DJ. 2022. Single Cell Transcriptomics of Atlantic Salmon (L.) Liver Reveals Cellular Heterogeneity and Immunological Responses to Challenge by. Front Immunol 13:984799. https://doi.org/10.3389/fimmu.2022.984799.

The Gene Ontology Consortium. 2018. The Gene Ontology Resource: 20 Years and Still GOing Strong. Nucleic Acids Res 47:D330–38.

Trapnell C. 2015. Defining Cell Types and States with Single-Cell Genomics. Genome Res 25:1491–8. https://doi.org/10.1101/gr.190595.115.

Wagner A, Regev A, Yosef N. 2016. Revealing the Vectors of Cellular Identity with Single-Cell Genomics. Nat Biotechnol 34:1145–60. https://doi.org/10.1038/nbt.3711.

Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, Chandak P, Liu S, Van Katwyk P, Deac A, et al. 2023. Scientific Discovery in the Age of Artificial Intelligence. Nature 620:47–60. https://doi.org/10.1038/s41586-023-06221-2.

Wang S, Pisco AO, Mcgeever A, Brbic M, Zitnik M, Darmanis S, Leskovec J, Karkanias J, Altman RB. 2021. Leveraging the Cell Ontology to Classify Unseen Cell Types. Nat Commun 12:5556. https://doi.org/10.1038/s41467-021-25725-x.

Windster JD, Sacchetti A, Schaaf GJ, Bindels EM, Hofstra RM, Wijnen RM, Sloots CE, Alves MM. 2023. A Combinatorial Panel for Flow Cytometry-Based Isolation of Enteric Nervous System Cells from Human Intestine. EMBO Rep 24:e55789. https://doi.org/10.15252/embr.202255789.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-Scale Single-Cell Gene Expression Data Analysis. Genome Biol 19:15. https://doi.org/10.1186/s13059-017-1382-0.

Xu C, Prete M, Webb S, Jardine L, Stewart BJ, Hoo R, He P, Meyer KB, Teichmann SA. 2023. Automatic Cell-Type Harmonization and Integration across Human Cell Atlas Datasets. Cell 186:5876–91.e20. https://doi.org/10.1016/j.cell.2023.11.026.

Yao Z, Van Velthoven CTJ, Kunst M, Zhang M, Mcmillen D, Lee C, Jung W, Goldy J, Abdelhak A, Aitken M, et al. 2023. A High-Resolution Transcriptomic and Spatial Atlas of Cell Types in the Whole Mouse Brain. Nature 624:317–32. https://doi.org/10.1038/s41586-023-06812-z.

Zeng H. 2022. What Is a Cell Type and How to Define It? Cell 185:2739–55. https://doi.org/10.1016/j.cell.2022.06.031.

Zhong J, Aires R, Tsissios G, Skoufa E, Brandt K, Sandoval-Guzmán T, Aztekin C. 2023. Multispecies Atlas Resolves an Axolotl Limb Development and Regeneration Paradox. Nat Commun 14:1–12. https://doi.org/10.1038/s41467-023-41944-w.