Artificial Intelligence Models for Cell Type and Subtype Identification Based on Single-Cell RNA Sequencing Data in Vision Science

Yeganeh Madadi, Aboozar Monavarfeshani, Hao Chen, W. Daniel Stamer, Robert W. Williams, and Siamak Yousefi

Abstract—Single-cell RNA sequencing (scRNA-seq) provides a high throughput, quantitative and unbiased framework for scientists in many research fields to identify and characterize cell types within heterogeneous cell populations from various tissues. However, scRNA-seq based identification of discrete cell-types is still labor intensive and depends on prior molecular knowledge. Artificial intelligence has provided faster, more accurate, and user-friendly approaches for cell-type identification. In this review, we discuss recent advances in cell-type identification methods using artificial intelligence techniques based on single-cell and single-nucleus RNA sequencing data in vision science. The main purpose of this review paper is to assist vision scientists not only to select suitable datasets for their problems, but also to be aware of the appropriate computational tools to perform their analysis. Developing novel methods for scRNA-seq data analysis remains to be addressed in future studies.

Index Terms—Artificial intelligence, review, single-cell RNA sequencing, vision science.

I. INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) technologies were established in 1990’s [1]. These technologies provide a powerful way for researchers to efficiently distinguish phenotypic heterogeneity of complex cell populations types in biological samples [2].

The utilization of scRNA-seq technology has profoundly transformed the way vision research is conducted. It has enabled large-scale and reliable characterization of various cell types in the ocular tissues resulting in a growing number of ocular (in particular, retinal) cell atlases from different species. These atlases have provided new insights into several rare and common eye diseases in two major ways. First, they have revealed cell-class and cell-type specific expression patterns for many disease-associated genes, and second, they have pinpointed similarities and differences among human and other animals’ ocular cell types validating their use in vision research.

Traditional scRNA-seq cell-type identification methods include two stages. First, cells are clustered through an unsupervised approach then clusters are annotated based on the canonical markers that are recognized in differentially expressed genes in each cluster [3], [4], [5]. In these methods, cluster annotation is very time-consuming because it requires broad literature knowledge about marker genes and forces calls whether to lump or split clusters based upon overlapping expression of marker genes.

In recent years, another class of single-cell data analysis models have been introduced to automatically identify clusters (e.g., cell types) without the requirement of manual annotation. These models utilize annotated scRNA-seq atlases for training the models and identify the clusters of unlabeled data based on the labels from the cells in the atlases. Most of these models use feature learning based on annotated training data (e.g., related reference atlases). Feature learning includes a set of techniques that allows a model to automatically discover the representations appropriate for classification task from raw data. These approaches have a common purpose, which is accurate cell annotation; however, they differ in proposed algorithms and the usage of prior knowledge [6], [7]. These approaches are useful when the user has access to large and well annotated datasets (atlases) [7], [8], [9].

The scRNAseq technology has been used in numerous research domains including vision and ophthalmology. For example, scRNA-seq was applied to retinal tissues for cell identification as early as 2015 [10]. Shortly after, several scRNA-seq approaches were applied to various visual system including...
retina [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], outflow pathways [29], [30], [31], cornea [31], [32], [33], [34], [35], iris [31], [34], [35], and ciliary body [31], [35], [36]. The eye can be used as a “window to the brain” as the structures of the eye are more accessible and can be more readily visualized using various imaging technologies compared to the brain proper. However, identifying eye’s cell types is highly challenging due to the specialized/hybrid nature of many cells that facilitate vision. Additionally, the numerous cell types of the eye collectively make it a structure of interest for investigating different health conditions. The light enters the eye and reaches the retina by passing through the cornea and lens, which in turn are responsible for concentrating the light appropriately on retina. The muscles of the iris, located posterior to the lens and the anterior cornea, regulates the angle of light that can enter the eye. The ciliary body and trabecular meshwork tissues control intraocular pressure via the regulation of the production and drainage of the aqueous humor, respectively. The retinal pigment epithelial (RPE) cells are located adjacent to photoreceptor cells lining the back of the orbit and serve to recycle visual pigment and provide nutrients to retinal photoreceptors. All these ocular tissues have major clinical impacts: For instance, lens opacification leads to cataracts, elevated intraocular pressure (IOP) may lead to glaucoma, and the retinal pigment epithelium dysfunction may eventually lead to age-related macular degeneration (AMD). Based on the WHO report, approximately 33.6 million people worldwide are blind in the 50 years and older age group [37]. The major distinguished causes of blindness are cataracts (∼15.2 million individuals), glaucoma (∼3.6 million individuals), and age-related macular degeneration (AMD) (∼1.8 million individuals) [38]. Over the past few years, significant scRNA-seq data have been generated from different eye tissues corresponding to these and other ocular conditions using various single cell type and subtype identification approaches. While important and promising, these findings would be made even more useful and valuable by benchmarking different artificial intelligence approaches to determine which method(s) is better suited to solve a particular problem or identify specific cell types. Additionally, a benchmarking review would aid investigators who are new to single cell and vision research choosing the most appropriate tools, thus avoiding potential confusion and the need to make selections based on trial and error.

In this article, we provide several taxonomies based on scRNA-seq datasets (atlases), well-established artificial intelligence approaches, and ocular diseases. First, we review the published scRNA-seq datasets related to the visual system. We then review the artificial intelligence approaches applied to broad cell type and subtype identification. We discuss the analysis pipeline and provide visualization of the scRNA-seq results. Furthermore, we review the identification of ocular diseases based on scRNA-seq data. Finally, we benchmark different artificial intelligence approaches based on three vision-related datasets, which are perfect for this type of analysis—not too big or small, and discuss advantages and disadvantages, as well as potential applications.

II THE scRNA-SEQ/snRNA-SEQ ATLASES (DATASETS) OF THE VISUAL SYSTEM

A. Atlases of the Mouse Retina

Macosko et al. [10] were among the first groups who applied high-throughput scRNA-seq (i.e., Drop-seq technology) and generated scRNA-seq data from ocular tissues. They used unsupervised machine learning methods and identified 39 cell clusters from 44808 single retinal cells. They annotated 33 cell types comprised of both rod and cone photoreceptors (PRs), retinal ganglion cells (RGCs), horizontal cells (HCs), bipolar cells (BCs), amacrine cells (ACs), Müller glia (MG), and non-neuronal cells. A follow-up study led by Shekhar [11] enriched mouse retinal bipolar cells (BCs) before subjecting them to the scRNA-seq. From the transcriptome of 23300 putative single BCs, they concluded a classification approach which identified 15 distinct BC subtypes including all types identified previously by Macosko et al. [10] and two additional novel BC subtypes. In total, they identified 18 BC clusters in this study. In another study, Rheame et al. [11] generated 6225 RGCs from mouse eyes, developed a computational pipeline, and analyzed scRNA-seq data. They identified 40 clusters of cells corresponding to 40 different subtypes of RGCs in which 30 RGC subtypes were recognized previously. They identified that most of the previously known RGC markers were expressed in multiple RGC clusters (subtypes) and only NPY, Jam2, Trhr, Pde1a, and Gna14 genes were expressed in a single cluster (RGC subtype). Therefore, they could determine the gene expression (variability) threshold required for RGC subtype segregation and diversification and presented a hierarchy from RGC cell types to subtypes. Further, they generated a portal for comparing gene expression in RGC subtypes. Tran et al. [14] in 2019 and Yan et al. [24] in 2020 adapted the pipeline introduced in Shekhar et al. [11] to classify a large number of RGCs and amacrine cells obtained from mouse retina. They identified 45 distinct clusters corresponding to RGC subtypes based on an scRNA-seq dataset with 35699 single cells and 63 subtypes of amacrine cells. Yan et al. [24] adopted a method of selective depletion based on 55287 single cells, of which 32523 cells were ultimately recognized as amacrine cells. They identified 63 clusters corresponding to different subtypes of amacrine cells which led to determination of class-specific marker for amacrine cells. They recognized molecular markers in each cluster and utilized them for the characterization of morphology in multiple types. They showed that the identified markers included all of the previously known AC type markers as well as numerous new markers. They observed most of the AC types expressed markers in canonical inhibitory neurotransmitters glycine or GABA, but some of the identified markers expressed neither in glycine nor in GABA. They also found transcriptomic relationships between AC types and recognized transcription factors that were expressed through single or multiple closely related cell types. Also, they observed that Tcf4 and Meis2 genes were expressed by most glycineric and most GABAergic types, respectively.

Heng et al. [15] proposed a scRNA-seq study to identify spontaneous uveoretinitis in Aire−/− mice by utilizing 64196 single
retinal cells. They analyzed the patterns of gene expression in various immune and retinal cells based on the abundance of different immune cell types and identified six different classes: Th1 cells, CD8a+ T cells, follicular helper T (Tfh) cells, regulatory T (Treg) cells, NK cells, and NK T cells. Sun et al. [24] constructed an atlas based on 14424 single cells collected from healthy mice and mice with diabetes to identify pathological alterations of diabetic retinopathy (DR). They determined several pathological variations of DR and introduced potential guides for DR therapy.

While the mouse visual system is an invaluable resource for studying neural circuit structure, function, and development, it has various shortcomings compared to the human visual system. Essentially, retinas, vascular arrangement, and lamina of primates, such as humans, differ from those of rodents but some other structures such as conventional outflow system are very similar. Particularly, unlike rodents, primates have fovea, a small retinal region located in the center of retina, that is primarily responsible for chromatic vision [40]. In order to address these shortcomings and gain a broad comprehensive understanding of specific mechanisms underlying cell heterogeneity in human vision retina, several groups have generated single cells from human and other primate visual systems Retina.

B. Retina Atlases in Non-Human and Human Primates

In 2019, Peng et al. [15] generated a retinal atlas profiling the foveal and peripheral retina of the macaque monkey separately. In order to enhance the probability of being able to identify rare RGC subtypes (e.g., ipRGC subtypes) localized within the highly rod-dominated peripheral region, they enriched RGCs using anti-Thy1 (from peripheral samples) and depleted rods using anti-CD73 in order to identify rare cell types and also subtypes. Based on 165000 single cells, they identified 64 foveal and 71 peripheral clusters corresponding to different cell types and subtypes. In several studies, macaque atlases have been used to annotate different clusters of retinal cells [17, 18, 19, 20, 21, 22, 23, 27]. Over the past few years, several research groups have extended related studies to foveal and peripheral retina of humans [19, 21, 22, 27].

Consistent with physiological and morphological studies, cell types in human and macaque retinas are highly similar. In 2019, Menon et al. [17] generated the first human retina single-cell transcriptomic atlas based on 20091 single cells. Their analysis was based on a multi-resolution network to identify different clusters corresponding to the retinal cell types and representative markers (e.g., genes expressed differentially). They found that human retinal microglia are more heterogeneous than previously thought and recognized some cell types (glia, cone photoreceptors, and vascular cells) associated with the risk of age-related macular degeneration (AMD). Lukowski et al. [18] compiled a transcriptome atlas from human retina by profiling 2009 single cells. They identified 18 transcriptionally distinct cell types representing all known neural retinal cells: retinal ganglion cells, Müller glia, bipolar cells, amacrine cells, rod photoreceptors, cone photoreceptors, horizontal cells, microglia, and astrocytes. Liang et al. [20] generated an atlas based on 5873 single nuclei collected from human retinal tissue. They identified major retinal cell types and their corresponding marker genes. More specifically, they identified several differentially expressed genes corresponding to photoreceptor cells that may serve as potential genes for inherited retinal diseases (IRD) and showed that their atlas may improve prioritization of genes related to human retinal diseases in comparison to mouse snRNA-seq and human bulk RNA-seq data.

Several teams have generated atlases of single cells collected from human fovea and peripheral retina [19, 21, 22, 23, 24, 25, 26, 27]. Voigt et al. [19] proposed an scRNA-seq for the human fovea and peripheral retina. They retrieved 8217 cells, including 3578 cells from fovea and 4639 cells from periphery. They found two clusters of cone photoreceptor cells based on differentially expressed genes. One cone cluster included 96% of foveal cells, while the other cluster included exclusively peripherally-isolated cone cells. The major marker for peripheral cones was Beta-Carotene Oxygenase 2. They observed a relative defect caused by this enzyme that may be due to the carotenoids accumulation which is primarily responsible for the yellow pigment in the macula.

The other study for understanding disease pathogenesis and identifying causal AMD genes was performed by Orozco et al. [23]. They presented an atlas of the human retina and retinal pigment epithelium (RPE) and comprised an expression quantitative trait loci (eQTL) atlas that included RPE/choroid and macula-specific retina, and single-nucleus RNA-seq from the human retina. They found enriched expression for AMD gene candidates in RPE cells. They identified 15 putative genes for AMD according to genetic association signals within AMD risk and eye eQTL, containing the genes TSPAN10 and TRPM1. Yan et al. [21] analyzed the transcriptomes of 84982 cells of the fovea and peripheral retina and recognized 58 clusters corresponding to different cell types and subtypes within retinal ganglion, horizontal, bipolar, photoreceptor, amacrine, and non-neuronal cells groups. Their follow-up analyses suggested that almost all cell types are common among two retinal regions, however there are significant differences in the gene expression level and the number of expressed genes among foveal and peripheral groups of shared cell types.

Cowan et al. [22] used human retinal tissues with multiple synaptic and nuclear layers for studying disease mechanisms and treatment in human retinas. They sequenced 285441 single cells from the fovea and periphery, choroid, and pigment epithelium of human retinas. Comparing the periphery to the fovea, they recognized regional characteristics and features of clusters and, through comparing organoid to organ, they determined that organoid cell type transcriptomes converge to those of peripheral retinas in adult humans.

Yi et al. [27] generated a transcriptomic atlas from 119520 single cells collected from the fovea and peripheral retina from humans and macaques of different ages. They observed that retinal molecular features were differed among the two species, indicating distinct regional and species dependent in human and macaque retinas. They also observed that aging in human retina occurred in specific cell types and regions, reflecting a foveal-to-peripheral gradient. Further quantification analysis of
scRNA-seq data showed that population of MYO9A− rods and horizontal cell subtypes significantly decreased during aging of retina.

C. Retina Atlases in Insects, Avians, and Fish

Ariss et al. [13] generated 11500 single eye disc cells including all main known cell types and determined the impact of an Rbf mutation within Drosophila eye development. They identified genes differentiating photoreceptor cells during axonogenesis and found a cell group that showed intracellular acidification primarily due to growth of glycolytic activity. Yamagata et al. [26] utilized scRNA-seq for generating a chick retina cell atlas to study the avian visual system. They recognized 136 cell types and 14 developmental or positional intermediates distributed between six classes: retinal ganglion, amacrine, photoreceptor, horizontal, bipolar, and glial cells. They adopted an approach based on CRISPR method, named eCHIKIN to annotate various clusters and to analyze selectively expressed genes to recognize the molecularly defined types of morphology. For Muller glia, they observed that transcriptionally different cells were regionally localized along the dorsal-ventral, anterior-posterior, and central-peripheral retinal axes. Additionally, they identified several cell types including immature horizontal cells, photoreceptors, and oligodendrocyte types which persist into late embryonic stages. Kolsch et al. [28] developed a systematic approach to classify RGCs in adult and larval zebrafish, then identified and validated antibody and in situ hybridization primarily due to growth of glycolytic activity. They recognized distinct trabecular meshwork and Schlemm’s canal cell populations and discovered additional distinct trabecular meshwork-Schlemm’s canal signaling pathways.

As light enters the eye, it passes through five layers of cornea including endothelium, epithelium, stroma, Bowman’s layer, and Descemet’s membrane. The cornea’s structure and transparency formations are maintained by various cell-types populated in each layer. Attempts to understand disease conditions and to regenerate corneal tissue requires extensive knowledge for modifying cell types across this heterogeneous tissue. Collin et al. [32] analyzed 21343 single cells collected from human corneas and adjacent conjunctivae and identified 21 clusters of cells. Subsequently, they extended their analysis to keratoconus disease and observed that activation of collagenase in corneal stroma and reduction of the number of limbal suprabasal cells could be key indicators for disease phenotypes. Català et al. [33] generated a single cell transcriptomic atlas based on 19472 single cells collected from corneal endothelium. They analyzed the corneal layers heterogeneity and identified HOMER3, CAV1, and CPVL expressions in the corneal epithelial limbal stem cell niche. They showed STMN1, CKS2, and UBE2C were individually expressed across highly proliferative transitional cells, NNMT was individually expressed through stromal kerocytes, and CXCL14 was expressed individually in suprabasal/superficial limbus. These results provide a basis for future amendments to current primary cell expansion protocols in order to enhance future profiling of corneal disease states. Wang et al. [34] introduced a transcriptomic atlas based on 16924 single cells collected from human corneal endothelium. The corneal endothelium is a major tissue for maintaining corneal clarity through mediating hydration by pump and barrier functions. Their findings provide novel insights into the development of Puchs endothelial corneal dystrophy (FEDC) and suggest that NEAT1 may offer an attractive method for treating FEDC.

The iris is responsible for controlling level of retinal illumination via regulation of pupil diameter. Wang et al. [35] presented a study that provided snRNA-seq data generated from iris cells of mice. They identified major cell types for the mouse iris and the ciliary body, which led to the detection of two kinds of iris stromal cells and also iris sphincter cells. They also characterized the diversities in cell-type transcriptomes within the dilated vs. resting states, then identified and validated antibody and in situ hybridization (ISH) probes to illustrate major iris clusters.

Youkilis et al. [36] introduced a transcriptomic atlas based on 10024 cells collected from ciliary body of mice and contiguous tissues, which play a major role in ocular homeostasis. They utilized scRNA-seq to assess the transcriptional signatures from the ciliary body and adjacent tissues. They identified two fibroblast signatures in the ciliary body cells (sclera and uvea), which were subsequently confirmed via in situ hybridization (ISH).
Van Zyl et al. [31] generated a single nucleus atlas of ocular anterior segment from the human eye. The anterior segment of human eye includes cornea, iris, crystalline lens, aqueous humor outflow pathways, and ciliary body. They profiled 195248 nuclei from anterior segment tissues and identified more than 60 cell types.

Table I shows the overview of scRNA-seq and snRNA-seq atlases generated based on vision system. It is worth mentioning that the Spectacle web tool houses all single cell atlases for vision [41].

III. ARTIFICIAL INTELLIGENCE APPROACHES IN CELL-TYPE IDENTIFICATION METHODS FOR scRNA-SEQ/snRNA-SEQ

Table II presents a taxonomy for artificial intelligence methods applied on scRNA-seq and snRNA-seq data to recognize cell-types and subtypes based on different machine learning approaches. In this taxonomy, we have classified models based on how they learn to find different clusters or classes. According to this issue, we have divided them into three main categories: unsupervised, semi-supervised, and supervised approaches. We have then provided sub-categories at each level and provided the references to those approaches.

A. Unsupervised Methods for Cell Type and Subtype Identification

Unsupervised cell-type identification methods recognize cell types and subtypes based on different clustering techniques. In unsupervised learning models, the data has no labels (cells have no type or subtype labels). Labeling the data means annotating the samples into different categories. Unsupervised learning models identify different clusters (groups) of samples based on (interesting) structures or patterns in the data. The labels will be propagated to the samples once the clusters were identified. Clusters (cell types or subtypes here) must be validated as there is no prior knowledge regarding the true labels. Unsupervised models can be further categorized to four subgroups including graph-based, partition-based, density-based, and deep neural network-based clustering. Some of these methods are discussed in more detail below.

Scmap [42] is a graph-based clustering technique that evaluates the maximum similarity between cells in the reference data and query data to identify different clusters corresponding to cell types. The optimum projection of a new query cell onto a reference data set is identified by the nearest neighbor approach. Seurat 3.0 [4] is a widely used toolkit for single-cell data analysis that employs this approach. Essentially, Seurat uses canonical correlation analysis (CCA) [3] and mutual nearest neighbors (MNNs) [60] approaches to identify cells with shared properties in the datasets. Seurat first jointly reduces the dimensionality of the reference and query datasets utilizing diagonalized CCA, then applies L2-normalization for the canonical correlation vectors. The model then searches MNNs in the common low-dimensional representation. Seurat encodes cellular relationships among datasets, which serve as the basis in all subsequent integration analyses.

In contrary to Seurat, Scmap learns cell-type-specific gene expression information from reference dataset only. As such, Scmap could be more vulnerable to the batch effect if reference and query datasets are generated in different batches. It is worth mentioning that although Seurat 3.0 employs the information (anchor pairs) extracted from the reference and query datasets, it doesn’t specifically use cluster information in the reference dataset. Many studies [10], [13], [15], [18], [19], [20], [25], [26], [27], [28], [29], [32], [33], [34], [35], [36] have used Seurat to perform clustering as part of their computational pipeline to identify cell types and subtypes. In the sCType [43] approach, the authors have introduced a marker database as well as a cell-type identification approach based on the Louvain algorithm for unsupervised cell-type annotation. This approach provides an end-to-end pipeline for identifying single cell types and subtypes based on a built-in marker database. It is worth mentioning that Yan et al. [21], [24] and Van Zyl et al. [30] performed unsupervised clustering approach based on Louvain algorithm with using Jaccard correction while Shekhar et al. [11], Tran et al. [14], and Cowan et al. [22] employed Infomap algorithm [61] in their graph-based unsupervised clustering approach.

Some of the approaches including Heng and colleagues [16] have used conventional unsupervised clustering models such as k-means. Based on this model, cell types and subtypes are identified by cross-referencing the clusters for expression of multiple known cell type-specific markers.

Rheaume et al. [12] used a density-based clustering approach, namely, DBSCAN [62] to identify cell types and subtypes. DBSCAN is an unsupervised clustering method that identifies distinctive clusters in data based on high adjacent density of samples, while separated from other clusters with low adjacent density of samples.

scScope [44] is a deep learning-based cell-type identification method, which utilizes a recurrent network layer for performing imputations iteratively on entries with zero-valued of input scRNA-seq data. scScope model lets imputed output to be iteratively amended through a selected number of recurrent steps. If number of recurrent steps is equal to 1, the model reduces to a standard autoencoder [63].

DESC [45] provides another unsupervised model based on deep learning algorithm which iteratively learns gene expression patterns corresponding to each cluster. Essentially, DESC first initializes different parameters obtained via an autoencoder then learns a nonlinear function for mapping the original scRNA-seq data in the high-dimensional space into the feature space in the low dimensional space based on optimizing the unsupervised clustering objective function in an iterative manner. By performing this iterative procedure, each cell moves to its nearest cluster centroid and balances technical and biological differences among clusters, and gradually decreases the influence of batch effect. ScAIDE [46] is another unsupervised deep learning clustering method that utilizes the autoencoder imputation network along with the distance-preserved embedding network (AIDE) for learning the data representations and then uses random projection hashing based on the k-means algorithm (RPH-kmeans) for accommodating the identification of rare cell types. There are several models based on neural networks and deep learning...
| Reference          | Year  |nof Cells|nof Clusters| Species | Tissues                            | Diseases     | Sequencing Instruments | Available Data File Formats | Deposited Data |
|--------------------|-------|---------|------------|---------|------------------------------------|--------------|------------------------|-----------------------------|-----------------|
| Macosko et al.     | 2015  |44,808   |39          | Mouse   | Retina                            | -            | Illumina NextSeq 500   | BAM, FASTA, GTF, TXT     | GSE63475        |
| Shekhar et al.     | 2016  |23,300   |18          | Mouse   | RGCs                             | -            | Illumina HiSeq 2500, Illumina NextSeq 500 | BAM | GSE81905        |
| Rheuma et al.      | 2018  |6822     |41          | Mouse   | RGCs                             | Glaucoma     | Illumina HiSeq 4000 | CSV | GSE115404       |
| Ariss et al.       | 2018  |11,500   |15          | Drosophila | Eye disc Retinoblastoma            | Illumina NextSeq 500 | TXT | GSE115476        |
| Tran et al.        | 2019  |35,699   |45          | Mouse   | RGCs                             | Glaucoma     | Illumina HiSeq 2500, Illumina NextSeq 500 | CSV | GSE137400       |
| Peng et al.        | 2019  |165,000  |Fovea:64 Periphere:71 | Macaque | F&PR                             | DR           | Illumina HiSeq 2500 | - | GSE118480       |
| Heng et al.        | 2019  |64,196   |12          | Mouse   | Retina                            | Uveoretinitis | Illumina HiSeq 2500, Illumina NovaSeq 6000 | MTX, TSV | GSE132229, GSE137537, GSE137847 |
| Menon et al.       | 2019  |20,091   |9           | Human   | Retina                            | AMD          | Illumina NextSeq 500 | MTX, TXT, TSV | GSE138056       |
| Lukowski et al.    | 2019  |20,009   |18          | Human   | Retina                            | -            | Illumina HiSeq 2500 | TXT, TSV | E-MTAB-7316     |
| Voigt et al.       | 2019  |8,217    |17          | Human   | F&PR                             | AMD          | Illumina HiSeq 4000 | TSV | GSE130636       |
| Liang et al.       | 2019  |5873     |7           | Human   | Retina                            | AMD          | Illumina HiSeq 2500 | TSV | GSE133707       |
| Yan et al.         | 2020  |84,982   |58          | Human   | F&PR                             | AMD, DR, Glaucoma | Illumina HiSeq 2500 | CSV | GSE148077       |
| Cowan et al.       | 2020  |285,441  |65          | Human   | F&PR, RPE, choroid Retina and RPE | AMD, Glaucoma | Illumina HiSeq 2500 | TXT | GSE148277       |
| Orozco et al.      | 2020  |100,055  |27          | Human   | Retina                            | AMD          | Illumina HiSeq 2500 | TSV | GSE133092, GSE135133  |
| Yan et al.         | 2020  |32,523   |63          | Mouse   | RACs                             | -            | Illumina HiSeq 2500 | CSV | GSE149715       |
| Patel et al.       | 2020  |8,758    |12          | Human   | OP                                | Glaucoma     | Illumina NextSeq 500 | - | PRJNA616025     |
| Van Zyl et al.     | 2020  |24,023   |19          | Human, Macaque, Pig, Mouse       | OP           | Glaucoma     | Illumina HiSeq 2500 | CSV | GSE146168       |
| Sun et al.         | 2021  |14,424   |28          | Mouse   | Retina                            | DR           | Illumina NovaSeq 6000 | XLS | GSE178121       |
| Yamagata et al.    | 2021  |40,000   |150         | Chick   | Retina                            | -            | Illumina HiSeq 2500 | CSV | GSE159107       |
| Yi et al.          | 2021  |119,520  |56          | Humans, Macaques | F&PR | AMD     | -            | XLSX | GSA-CRA002680, GSA-HRA00182 |
| Ko’lsch et al.     | 2021  |32,679   |32          | Zebrafish | RGCs | -            | Illumina HiSeq 2500 | CSV | GSE152842       |
| Collin et al.      | 2021  |21,343   |21          | Human   | Cornea                            | Retinoblastoma | Illumina NovaSeq 6000 | CSV, TXT | GSE155683       |
| Català et al.      | 2021  |19,472   |15          | Human   | Cornea                            | -            | Illumina NovaSeq 6000 | MTX, TSV | GSE186433       |
| Wang et al.        | 2021  |34,357   |10          | Mouse   | Iris and ciliary body Iris disorders | -            | Illumina NovaSeq 6000 | MTX, TSV | GSE183690       |
| Youkalis et al.    | 2021  |10,024   |22          | Mouse   | CB&CT                             | -            | Illumina NovaSeq 6000 | MTX, TSV | GSE178667       |
| Thomson et al.     | 2021  |26,008   |25          | Mouse   | OP                                | Glaucoma     | Illumina HiSeq 4000 | CSV | GSE168200       |
| Wang et al.        | 2022  |16,924   |4           | Human   | Corneal endothelium FECD          | Illumina No-vaSeq 6000 | - | GSA-HRA00781 |
| Van Zyl et al.     | 2022  |195,248  |60          | Human   | Anterior segment                  | Glaucoma     | Illumina HiSeq 2500, Illumina NovaSeq 6000 | CSV | GSE199013       |

RGCs = Retinal Ganglion Cells, F&PR = Fovea and peripheral retina, RPE = Retinal pigment epithelium, CB&CT = Ciliary body and contiguous tissues, RACs = Retinal amacrine cells, OP = Outflow pathways, DR = Diabetic Retinopathy, FECD = Fuchs endothelial corneal dystrophy.
TABLE II
| TAXONOMY OF CELL-TYPE IDENTIFICATION METHODS THAT WERE INCLUDED IN THIS STUDY |
|---------------------------------|-------------------------------|
| Unsupervised                    | Scmap [42], Neurat [4]        |
| Graph-based                     | [15], [11], [13], [14], [15], |
| Nearest neighbor                | [17], [18], [19], [20], [22], |
| [29], [25], [26], [27], [28],  |
| [32], [33], [21], [23], [24],  |
| [35], [36], [34], [35]          |
| scType [43]                     |
| Partition-based K-means clustering | [16]                        |
| Density-based DBSCAN clustering | [12]                         |
| Deep NN-based                   | Recurrent network ScScope [44]|
| Autoencoder                     | DESC [45], SCAIDE [46],       |
| scETM [47]                      | Hierarchical Bayesian scVI [48]|
| Semi-supervised                | Deep NN-based Autoencoder     |
| Similarity-based                | DESC [49], ScDCC [50]        |
| General classifier-based        | XGBoost CaSTLe [52]           |
| Logistic regression SCAFE [53]  |
| Fisher’s linear discriminant analysis ScID [54] |
| Capsule networks ScCapsNet [55] |
| Deep NN-based                   | Weighted GNN ScDeepSort [56]  |
| Transfer learning-based         | Hierarchical FFNN NeCoA [57] |
| ScLearn [51] [52]               |
| scLearn [51]                    |
| scLearn [51]                    |

Semi-supervised learning approaches are promising when a few labeled data are available thus allowing unsupervised components to complement their training with unlabeled data [65].

SSL approaches may address dropout problems in single cell data. Dropout is essentially one of the main challenges in scRNA-seq data analysis. Dropouts, the extra false zero expressions, causes a skewed distribution of gene expressions and result in the misclassification of cell types [66]. Recent advances in combinatorial indexing-based or droplet-based sequencing methods have significantly grown the throughput from thousands to more than a million cells for a single experiment and caused more intense dropout problems because of shallow sequencing depth in per cell [67]. In the following, we will discuss a few SSL-based cell identification methods.

DISC [49] is an autoencoder-based semi-supervised deep learning network method to infer gene structure and expression of single cell data generated based on the dropout technology. Semi-supervised learning here generates a reliable imputation approach via learning information from zero-count genes and positive genes, that can be treated as unlabeled and labeled data, respectively. DISC works well in scenarios that the information from the labeled data is limited. Single Cell Deep Constrained Clustering (scDCC) [50] is another semi-supervised clustering approach based on autoencoder that encodes prior knowledge in constraint information, that is integrated into the clustering algorithm through a loss function. Also, scDCC integrates domain knowledge for the clustering step.

C. Supervised Methods for Cell-Type and Subtype Identification

Different approaches in this category typically utilize a previously annotated dataset as a reference to train supervised machine learning classifiers and use the trained model to identify cell types and subtypes from unlabeled data. However, in such supervised methods, it is expected that the reference and query datasets resemble each other, which is not often the case. This poses challenges in successful label propagation [60].

scLearn [51] is a supervised single cell classification method that utilizes the similarity among single cells in the query dataset and cluster centroids of cells in the reference dataset based on information learned from reference datasets. scLearn can identify novel cell types not presenting in the reference datasets. It uses a multilabel single-cell assignment approach to assign a single cell to proper time status and cell type simultaneously.

General classifier-based methods are another subgroup of supervised cell type identification methods. CaSTLe [52], classification of single cells through transfer learning, is a supervised method based on the XGBoost classifier [68]. This method selects genes with mutual information gain and genes with top mean expression, then removes correlated genes before classification. All these steps are performed to ensure the reference and query datasets are brought into a common denominator to perform an accurate transfer of the classification model. SCAFE [53] presents a self-projection method based on linear regression [69], and ScID (Single Cell IDenification) [54] utilizes the Fisher’s Linear Discriminant Analysis
to recognize transcriptionally related cell types among scRNA-seq datasets. This method extracts markers through the reference dataset and then weighs their relevance on the target dataset via learning a classifier based on the putative population of different cells either expressing or not expressing these genes.

scCapsNet [55] and scDeepSort [56] are a subgroup of supervised deep neural network-based methods in proposed taxonomy. scCapsNet [55] has an interpretable deep-learning architecture utilizing capsule networks (scCapsNet). Capsule structure is a neuron vector that represents the properties of a specific object and captures hierarchical relations. scCapsNet creates the decision-making black box transparent via analyzing internal weight parameters between capsule structures. scDeepSort [56] is a pre-trained cell-type identification approach based on weighted graph neural networks (GNN) model. GNN is one of the popularly-used deep learning methods [71] that captures graph dependency via message passing among the graph nodes and keeps a state which represents required information from its neighborhood with ideal depth [72]. NeuCA [57] computes the mean gene expression of each cell group then evaluates the correlation matrix among different cell types. If cell types are highly correlated, the model constructs a tree structure based on hierarchical feed-forward clustering and subsequently trains several neural networks (based on the tree structure) to perform the final classification task.

ItClust [58] has received more attention compared to unsupervised models as it utilizes the information from external well-annotated reference datasets. As the performance of supervised methods are highly dependent on the quality of the reference dataset, ItClust employs a transfer learning approach [73] that takes advantage of cell type-specific gene expression information learned from a high-quality reference dataset, to help classify cell types on a newly generated query dataset. The approach considers highly variable genes within the reference data in order to extract relevant information. This ensures that transferred expression patterns will be useful for dividing cell types based on the new datasets. SCTL [59] is another deep transfer learning method to detect cell types and also subtypes in the scRNA-seq data. SCTL includes four loss functions where one loss function represents domain adaptation, and three loss functions represent the adversarial network. These loss functions eventually minimize the classification error.

IV. ANALYSIS PIPELINE AND VISUALIZING scRNA-SEQ/snRNA-SEQ RESULTS

In this section, we discuss key steps in scRNA-seq data visualization. Numerous statistical and machine learning methods are applied to the gene expression data to gain insight into the transcriptome heterogeneity of cell types. These techniques typically include normalization, scaling, dimension reduction, clustering, and classification. However, visualization plays an important role in both qualitative and quantitative scRNA-seq data analysis (see Fig. 1).

There are several different techniques for visualizing scRNA-seq/snRNA-seq results (Fig. 1(a)–(f)). The t-distributed stochastic neighbor embedding (t-SNE) is a widely used statistical method to visualize high-dimensional data by assigning each data point to a location on a low-dimensional (i.e., two-dimensional) map. Fig. 1(a) shows clusters of the molecular diversity for cells using tSNE on a two-dimensional map [74]. Each point corresponds to a single cell, with transcriptionally similar cells being clustered together. Cells are typically colored differently based on their defined cluster identity. Fig. 1(b) shows the feature plot of cells in panel (A) at which cells are colored relying on their level of expression of genes of interest (i.e., genes that were differentially expressed between clusters). Uniform manifold approximation and projection (UMAP) is similar to t-SNE. Both t-SNE and UMAP build a graph which represents data into high dimensional space and reconstruct the graph into a lower dimensional space. Fig. 1(c) shows violin plot of the expression levels (y axis) of single cells in Panel A based on a specific gene (i.e., a gene that was differentially expressed in cluster number 4 compared to other clusters). Each violin represents the distribution of expression levels of cells within that particular cluster. The box and whisker plots within each violin correspond to the median value (black horizontal line), range (vertical lines), interquartile range (bars), and outliers (dots). Fig. 1(d) shows a sample dot plot visualizing the expression pattern (both number
of cells expressing a gene and the level of expression) of a set of genes across different clusters (x-axis). The size of each circle indicates the percentage of cells in that cluster expressing the gene, and color saturation represents normalized expression level. Along with the dot-plot, a dendrogram calculated using hierarchical clustering on top typically represents transcriptional interrelationships among clusters. Fig. 1(e) shows a sample heat map plot. This ordered heatmap plot represents the expression patterns of specific marker genes for single cells grouped by their cells in different clusters. Fig. 1(f) shows a sample confusion matrix plot representing transcriptional correspondence between clusters in two different datasets. Colors indicate the proportion of cells of a given cluster in one dataset assigned to a corresponding cluster in the other dataset in which the classification algorithm trained on cells in one of the datasets.

V. IDENTIFICATIONS OF OCULAR DISEASES BASED ON scRNA-SEQ/sRNA-SEQ DATA

In this section, we discuss ocular diseases and related tissues which have been studied on single-cell and single-nucleus RNA sequencing data. Most irreversible vision losses are caused by retinal diseases that explain the focus of most research on retina. The three prevalent and irreversible leading causes of vision loss are diabetic retinopathy, age-related macular degeneration (AMD), and glaucoma. Based on a single cell study, Lukowski et al. [18] observed that MALAT1, a long non-coding RNA, plays a major role in retinal homeostasis and disease [75] and they recognized that MIO-M1 cells have high levels of the thymosin beta 4 gene (TMSB4X) that has been related to glioma malignancy [76], and the calycin gene (S100A6), that is implicated in cone-associated or macular diseases [77]. These results show the differences and similarities in MIO-M1 for human retinal glial cells. Many groups have utilized scRNA-seq/snRNA-seq data to assess retinal cell types which express genes implicated in different retinal diseases [15], [21], [22], [23], [27]. For example, most genes which are mutated in retinitis pigmentosa and initially affect rods, are expressed via rods or via retinal pigment epithelial cells and required for rod viability. Also, most genes mutated in autosomal dominant optic atrophy and Hereditary optic neuromatrophies, that result in RGC death, are selectively expressed in RGCs; and many susceptibility genes related to diabetic retinopathy are expressed in vascular endothelial cells. There are various studies on AMD pathogenesis. CFH and ARMS2/HTRA1 genes impart biggest risk in AMD disease [78], Voigt et al. [19] evaluated the differentially expressed genes among cells of foveal versus peripheral origin. Menon et al. [17] identified various cell types associated with AMD. Their results suggest that the genetic risk variants related to AMD affect cone photoreceptors, and they emphasized the importance of vascular and glial cells for disease pathogenesis. They found that expression of COL4A3, HTRA1, and vascular endothelial growth factor (VEGFA) had high scores for leading to AMD. Orozco et al. [22] identified enriched expression of AMD candidate genes for RPE cells. They identified TSPAN10 and TRPM1 which were enriched in retinal pigment epithelium, as causal genes that have high impact in early AMD disease. Yi et al. [27] reported that genes related to AMD are highly enriched in cones and foveal MG, suggesting a relation of regional cell subtype with this disease. Lu et al. [79] showed that compositional changes are more pronounced in the macula in rods, endothelium, microglia, astrocytes, and Müller glia in the transition from normal to advanced AMD. They also identified enrichment in coagulation and complement pathways, signaling pathways, tissue remodeling, and antigen presentation, including PI3K-Akt, Rap1, Toll-like, and NOD-like. Sun et al. [25] studied diabetic retinopathy and identified four stress-inducible genes Rmb3, Cirrb, Mt1, and Mt2 which commonly exist in most retinal cell types. Diabetes increases the inflammatory factor gene expressions in retinal microglia and stimulates the immediate early gene expressions (IEGs) in retinal astrocytes. Van Zyl et al. [30] studied glaucoma cases and recognized the cell types which represent gene expressions implicated in glaucoma. They found that several genes, such as MYOC, PITH2, CYP1B1, Cav1, and Cav2 are implicated in glaucoma. In their other study [31], they showed MYOC, ANGPT1, LMX1B, ANGPT2, PITH2, LTB2, FOXC1, and CPAMD8 were associated with glaucoma. Patel et al. [29] studied assigning glaucoma-relevant genes to outflow cell clusters. They found that MYOC, PDPN, ANGPTL7, CHI3L1, and ANGPT1 were highly expressed in TM1 and TM2 cell types, while CAV1, CAV2, Tie2 (TEK), NOS3, ANGPT2, and PLAT, were highly expressed in vascular endothelial and lymphatic-like cell types.

In Fig. 2, we show the percentage of scRNA-seq/snRNA-seq studies based on different ocular diseases and ocular tissues from different species We found that in vision and ophthalmology, mouse, human, macaque, and chick were studied the most (55%, 19%, 11%, 9%, respectively), followed by drosophila, pig, and zebrafish that were studied similarly (about 2% each).

From mouse studies, about 89%, 7%, 4% are focused on retina, iris, and outflow pathways tissues, respectively. It is worth mentioning that almost all of the studies on pig have only been applied to the outflow pathways tissues. From all the studies performed on macaque, about 82% are on retina and 18% on outflow pathways tissues. Almost all zebrafish studies have only been conducted on retina. From studies done on human tissues, about 90% have analyzed retina and the remaining have focused on the outflow pathways tissues, respectively. From the studies performed on chick, about 22% were performed on retina and 78%, were applied on cornea (see Fig. 2(a)).

Also, we observed that most single cell studies have focused on AMD, diabetic retinopathy, glaucoma, and uveitis. About 33% of the single cell studies were focused on glaucoma, whereas AMD, diabetic retinopathy, uveitis, and other uvea diseases constituted 29%, 17%, 4% and 17% of studies, respectively (see Fig. 2(b)).

VI. BENCHMARKING CELL-TYPE IDENTIFICATION METHODS FOR scRNA-SEQ/sRNA-SEQ DATA

A. BenchmarkDatasets

Benchmarking could aid investigators who are new to single cell and vision research to select the most appropriate tools...
Fig. 2. Percentage of single cell studies in vision and ophthalmology based on different species, tissues, and diseases. (a) Percentage of studies conducted on mouse, pig, macaque, zebrafish, human, chick, drosophila were 55%, 2%, 11%, 2%, 19%, 9%, and 2%, respectively. From studies done on mouse species, about 89%, 7%, 4% are on retina, iris, and outflow pathways tissues, respectively. Almost all pig studies have only been done on outflow pathways tissues. From studies performed on macaque, 82% were on retina and 18% on outflow pathways tissues. Almost all zebrafish and drosophila studies have only been done on retina. From studies conducted on human, about 90% are on retina while about 10% are focused on outflow pathways tissues. From studies on chick, about 22% are focused on retina while nearly 78% are focused on cornea. (b) The percentage of studies performed on different diseases include about 33% on glaucoma, 29% on age related macular degeneration (AMD), 17% on diabetic retinopathy (DR), 4% on uveoretinitis, and 17% on other uvea diseases.

levels of challenges in identification of each cell type (see Table III).

The first dataset, selected from the Macosko et al. [10] study, included 44808 single cells of 39 various retinal cell types (and subtypes) and 24658 common genes using Droplet-based RNA-seq technology. Single cells in this dataset were obtained from retinas of several 14-day-old wild-type C57BL/6 mice. As these mice, within each strain are genetically identical, they highly present uniform inherited characteristics and response to experimental treatments. Experimental workflow of this study was performed by using seven various batches generating 3226, 6020, 8336, 5683, 6991, 6971, and 7581 cells (total 44808 cells).

The second dataset, selected from the Rheaume et al. [12] study, included 6225 single RGCs comprising 41 different RGC subtypes and 13616 common genes generated using a single batch based on Droplet-based RNA-seq technology. Single RGCs were obtained from retinas of eight postnatal C57Bl/6 mice at day five including both sexes.

The third dataset, acquired from the Tran et al. [14] study comprised of 35699 single RGCs in 45 various subtypes and 18222 common genes. Single RGCs were collected from retinas of adult C57BL/6 mice with ages ranged from 6 to 20 weeks. Single RGCs were generated in three batches using the droplet-based RNA-seq technology. Characteristics of the cells and cell types are shown in Table III.

We randomly split the cells in these datasets into two subsets, where 80% of cells were allocated to the training data and the remaining 20% of cells were allocated to testing data. The performance of the single cell-type identification methods was assessed by cross-validation (calculating the mean accuracy based on five rounds of testing).

B. Data Preprocessing

Single cell datasets, count matrices, and annotated labels were downloaded from the public resources (GSE63473, GSE115404, GSE137400). We first applied the Seurat [4] R package (version 4.0.6) to normalize and scale the transcriptome datasets. We applied two exclusion criteria: we excluded cells with fewer than 200 expressed genes then removed genes which were expressed in fewer than three cells. We then selected a subset of highly variable genes to further reduce dimensionality. Specifically, we computed the dispersion (variance/mean) and the mean expression levels of each gene [10] and computed the z-normalized dispersion level of all genes to find highly variable genes. We used the same cutoff level of 1.7 as was used in the Macosko et al. [10], which is also the default parameter in Seurat package. The preprocessing led to inclusion of about 2000 highly variable genes for the downstream analysis.

C. Benchmark Methods

Among the forty-five methods discussed in this review, we included seventeen published cell identification methods that their source codes or program was publicly accessible (see Table IV).

These methods are divided into three main categories: (1) Unsupervised cell-type identification methods, which are based...
TABLE III

| Macosko et al. [10]               | Rheuma et al. [12]            | Tran et al. [14]          |
|-----------------------------------|-------------------------------|---------------------------|
| Cluster name                      | Cells                         | Cluster name              | Cells                         |
| HC                                | 252                           | P5 RGC0                   | 161                           | W3-1ike1: RGC1                   | 3,000                         |
| RGC                               | 432                           | W3D1: P5                  | 426                           | W3D1: RGC2                       | 2,859                         |
| AC                                | 289                           | P5 RGC2                   | 188                           | F-mnti-ON: RGC3                   | 1,990                         |
| AC                                | 73                            | P5 RGC3                   | 136                           | F-mnti-OFF: RGC4                   | 1,868                         |
| AC                                | 77                            | F-mnti-ON: P5             | 329                           | J-RGC: RGC5                       | 1,715                         |
| AC                                | 211                           | P5 RGC5                   | 66                            | W3B8: RGC6                       | 1,590                         |
| AC                                | 326                           | P5 RGC6                   | 52                            | RGC7                            | 1,579                         |
| AC                                | 159                           | F-mnti-OFF: P5            | 268                           | RGC8                            | 1,258                         |
| AC                                | 350                           | F-RGC: P5                 | 95                            | T-RGC: RGC9                       | 1,223                         |
| AC                                | 191                           | P5 RGC9                   | 185                           | RGC10                           | 1,170                         |
| AC                                | 214                           | P5 RGC11                  | 143                           | RGC11                           | 990                           |
| AC                                | 274                           | P5 RGC11                  | 88                            | oedSGC-N: RGC12                   | 953                           |
| AC                                | 50                            | P5 RGC12                  | 135                           | W3-1ike2: RGC13                   | 943                           |
| AC                                | 111                           | W3-1ike1: P5              | 429                           | RGC14                           | 879                           |
| AC                                | 73                            | J-RGC: P5                 | 235                           | RGC15                           | 865                           |
| AC                                | 262                           | T-RGC-S2: P5              | 91                            | oedSGC-DV: RGC6                   | 829                           |
| AC                                | 375                           | P5 RGC16                  | 135                           | T-RGC-S1: RGC17                   | 828                           |
| AC                                | 83                            | P5 RGC17                  | 121                           | RGC18                           | 826                           |
| AC                                | 127                           | W3D3: P5                  | 80                            | RGC19                           | 779                           |
| AC                                | 389                           | P5 RGC19                  | 115                           | RGC20                           | 711                           |
| AC                                | 254                           | P5 RGC20                  | 224                           | T-RGC-S2: RGC21                   | 687                           |
| AC                                | 274                           | P5 RGC21                  | 102                           | MX: RGC22                        | 610                           |
| AC                                | 264                           | P5 RGC22                  | 48                            | W3D2: RGC23                      | 601                           |
| Rods                              | 29,400                        | P5 RGC23                  | 93                            | RGC24                           | 553                           |
| Cones                             | 1,868                         | P5 RGC24                  | 89                            | RGC25                           | 542                           |
| BC                                | 2,217                         | P5 RGC25                  | 175                           | RGC26                           | 534                           |
| BC                                | 664                           | P5 RGC26                  | 233                           | RGC27                           | 529                           |
| BC                                | 496                           | W3-1ike2: P5              | 147                           | F-mnti-OFF: RGC28                 | 517                           |
| BC                                | 591                           | P5 RGC28                  | 100                           | RGC29                           | 499                           |
| BC                                | 636                           | P5 RGC29                  | 124                           | W3D3: RGC30                      | 491                           |
| BC                                | 512                           | P5 RGC30                  | 186                           | M2: RGC31                        | 444                           |
| BC                                | 320                           | N: P5                     | 168                           | F-RGC: RGC32                     | 407                           |
| BC                                | 849                           | P5 RGC32                  | 133                           | M1a: RGC33                       | 332                           |
| BC                                | 1,624                         | P5 RGC33                  | 108                           | RGC34                           | 312                           |
| A                                 | 54                            | W3B8: P5                  | 155                           | RGC38                           | 310                           |
| F                                 | 85                            | P5 RGC35                  | 70                            | RGC38                           | 236                           |
| V                                 | 252                           | P5 RGC36                  | 183                           | RGC37                           | 213                           |
| P                                 | 63                            | P5 RGC37                  | 135                           | F-mnti-ON: RGC38                  | 207                           |
| M                                 | 67                            | P5 RGC38                  | 150                           | RGC39                           | 202                           |
| M                                 | 39                            | P5 RGC39                  | 44                            | M1b: RGC40                       | 174                           |
| M                                 | 40                            | P5 RGC40                  | 20                            | sON-T: RGC41                      | 136                           |
| M                                 | 43                            | P5 RGC43                  | 44                            | sOFF-S: RGC42                     | 113                           |
| M                                 | 40                            | P5 RGC40                  | 20                            | sON-S4: RGC43                     | 106                           |
| M                                 | 62                            | P5 RGC44                  | 62                            | aOFF-T: RGC45                     | 54                            |

HC = Horizontal Cells, RGC = Retinal Ganglion Cells, AC = Amacrine Cells, BC = Bipolar Cells, M = Muller Glia, A = Astrocytes, F = Fibroblasts, V = Vascular Endothelium, P = Pericytes, M = Microglia.

TABLE IV

| Method           | Availability                  | Code Ref. Year | Category          |
|------------------|-------------------------------|----------------|-------------------|
| DESC             | https://github.com/Pleuzer/desc| Python [45]    | 2020              | Unsupervised      |
| scVI             | https://github.com/VosefLAB/sclvi| Python [46]   | 2018              | Unsupervised      |
| Smap             | https://github.com/benmberg/lab_smap| R [42]      | 2018              | Unsupervised      |
| Seurat           | https://github.com/catialab/seurat| R [4]        | 2019              | Unsupervised      |
| SCAIDE           | https://github.com/tlzlabs/caIDE| Python [46]    | 2020              | Unsupervised      |
| scETM            | https://github.com/bu2003/scETM| Python [47]    | 2021              | Unsupervised      |
| DISC             | https://github.com/sie-lab/DISC| Python [49]    | 2020              | Unsupervised      |
| ScDCC            | https://github.com/georgiosScDCC| Python [50]    | 2021              | Semi-supervised   |
| SCAF             | https://github.com/SCAFscf/scf| Python [51]    | 2021              | Supervised        |
| liCluster        | https://github.com/janhuspen/lt_cluster| Python [58] | 2020              | Supervised        |
| ScDeepSort       | https://github.com/STUFS- Lab/ScDeepSort| R [56]      | 2021              | Supervised        |
| ScLearn          | https://github.com/bm2-lab/ScLearn| R [51]        | 2020              | Supervised        |
| ScCapsNet        | https://github.com/wangl/ScCapsNet| Python [55] | 2020              | Supervised        |
| ScID             | https://github.com/BataSalab/ScID| Python [54]    | 2020              | Supervised        |
| CaSTIte          | https://github.com/wulab/CaSTIte| R [52]        | 2018              | Supervised        |
| NeuCA            | https://github.com/biocca/NeuCA| Python [57]    | 2022              | Supervised        |
| ScTL             | https://github.com/Panlab/ScTL| Python [59]    | 2022              | Supervised        |

D. Evaluation Indicators

We evaluated the performance based on accuracy, speed, and memory usage metrics. Accuracy is defined as the ratio of cells that are correctly predicted cells, divided by the total number of annotated cells [9]. The memory requirements were obtained via reading rss (resident set size) attribute (which is returned through calling Process().memory_info() of psutil in the Python package) [80]. For the other approaches based on R language, we used the reticulate package [81] for calling the mentioned Python function for having consistency.

E. Results

1) Accuracy of Cell Type Identification: We compared seventeen cell-type identification methods for correctly identifying the number of cell types through applying each method on three ocular datasets [10], [12], [14] that contain 39, 41, and 45 cell on unsupervised clustering of the transcriptomes [4], [42], [45], [46], [47], [48]. (2) Semi-supervised cell-type identification methods, which require some limited labeled data in order to allow models to complement their training with unlabeled data [49], [50]. (3) Supervised cell-type identification approaches, which need a training dataset labeled with the corresponding cell populations for training the classifier [51], [52], [53], [54], [55], [56], [57], [58], [59].
Fig. 3. Accuracy (%) of seventeen cell-type identification methods based on (a) macosko et al. (b) Rheaume et al. (c) Tran et al. benchmark ocular datasets.

2) Accuracy of Batch Correction: We compared seventeen cell-type identification methods and evaluated the accuracy of the models when data has been generated based on several different batches using the Macosko et al. dataset. From seven different batches of this dataset, we assigned one single batch as testing dataset and the rest of the other batches as training dataset and reaped this process seven times to cover all seven batches as testing. Substantially smaller number of single cells exist in Batch B1 compared to the B2-B7 batches. Fig. 4(a)–(g) shows the accuracy (and standard error) of these methods based on batches B1 to B7, respectively. The benchmarking results on Macosko et al. dataset showed that SCTL and ItClust generate the highest accuracy in comparison to the other methods. Both methods use a transfer learning-based algorithm, which improves cell-type identification particularly on datasets with varying number of cells from different cell types and removes the batch effect better than the other methods. In benchmarking the methods based on the Rheaume et al. and Tran et al. datasets, with over 40 cell types, NeuCA, ItClust, and SCTL outperformed the other methods. Generally, the results on all three datasets show that supervised cell identification methods obtained higher accuracy. In average, these three methods, ItClust, NeuCA, and SCTL, obtained the highest accuracy based on all three datasets.

3) Running Time and Peak Memory Usage: We assessed the computational time (Fig. 5(a)) and peak memory usage (Fig. 5(b)) for cell-type identification methods based on different number of cells (sample sizes) randomly selected from the Macosko et al. dataset.

The experiments were performed on a workstation with 16 CPU cores at 3.90GHz, 64GB of memory and one GEFORCE RTX 3070 graphical processing unit (GPU) card.
As expected, almost all of the neural network-based methods (for example, ScDeepSort, ScCapsNet) need more computational time than the other approaches, since a large number of parameters are required to be optimized. Generally, unsupervised cell-type identification models show excellent performance in terms of speed and almost all of these methods require no more than 5 minutes to complete cell type or subtype identification.

To compare the computational time of the cell-type identification methods and to see how they are scaled when the number of cells increases, we randomly selected subsets of cell from the Macosko et al. dataset and benchmarked the methods. Results showed that the increase in computational time of Scmap and ScAIDE methods is smooth, suggesting that computational time increases almost linearly with increasing the sample size. Fig. 5(b) shows the performance of cell-type identification methods based on computational time.

To compare the peak memory usage of the cell-type identification approaches, we randomly selected subsets of cells from the Macosko et al. dataset and compared the methods. Results indicated that the scVI, scETM, and DESC required least amount of memory to complete the analyses.

Collectively, results suggest that applying cell-type identification methods on large datasets require larger memory sizes and longer run time thus the optimum size and computational time may be optimized based on the estimated numbers in Fig. 5.

VII. DISCUSSION

We discussed numerous vision-related scRNA-seq datasets and explained overall advantages and disadvantages of different cell-type identification models. We showed that scRNA-seq data analysis methods based on supervised learning performed better than methods based on unsupervised or semi-supervised models in terms of accuracy.

Transfer learning-based (e.g., IClust, SCTL) and Deep NN-based (e.g., NeuCA) methods, in comparison, performed favorably in general. Therefore, in scenarios in which we deal with a large number of labeled samples, supervised models may generate higher accuracy in identifying cell types, however if dealing with unlabeled data, unsupervised methods (based on Autoencoder, e.g., scETM) may perform better.

In terms of computational complexity, almost all of the neural network-based approaches (e.g., ScDeepSort, and ScCapsNet) require more computational time than the other methods since a large number of parameters are required to be optimized. Generally, unsupervised cell-type identification approaches show excellent performance in terms of speed.

In terms of peak memory usage, unsupervised methods use less memory. In fact, Hierarchical Bayesian (e.g., scVI) and Autoencoder (e.g., scETM, and DESC) methods required the least amount of memory to complete the analyses.

VIII. CONCLUSION

In this study, we reviewed different single cell studies conducted in in vision science. We described numerous vision-related scRNA-seq/snRNA-seq atlases and provided taxonomies based on different species, tissue, and ocular diseases. We also described and benchmarked different machine-learning-based models in single cell data analysis and categorized models into three main subgroups: unsupervised, semi-supervised, and supervised cell-type and subtype identification methods. We also provided insights to scRNA-seq data visualization and provided several approaches for effectively visualizing the outcome in single cell studies. Finally, we benchmarked seventeen single cell type and subtype identification methods based on three different datasets encompassing diverse single retinal cell types and subtypes. We provided metrics on the accuracy, computational time, and memory usage thus allowing vision researchers to optimize their available computational resources when working with single cell transcriptomic datasets. Our study provides a valuable review on available vision-related single cell datasets and scRNA-seq data analysis techniques and discusses future development of scRNA-seq based cell-type identification methods in vision science.

REFERENCES

[1] J. Eberwine et al., “Analysis of gene expression in single live neurons,” Proc. Nat. Acad. Sci., vol. 89, no. 7, pp. 3010–3014, 1992.
[46] K. Xie, Y. Huang, F. Zeng, Z. Liu, and T. Chen, “scAIrDE: Clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types,” *NAR Genomic. Bioinf.*, vol. 2, no. 4, Dec. 2020, Art. no.lqa082, doi: 10.1093/nargab/lqa082.

[47] Y. Zhao, H. Cai, T. Zhang, J. Tang, and Y. Li, “Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data,” *Nature Commun.*, vol. 12, no. 1, Sep. 2021, Art. no. 5261, doi: 10.1038/s41467-021-25534-2.

[48] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, “Deep generative modeling for single-cell transcriptomics,” *Nature Methods*, vol. 15, no. 2, pp. 1053–1058, Dec. 2018, doi: 10.1038/s41592-018-0029-2.

[49] Y. He, H. Yuan, C. Wu, and Z. Xie, “DISC: A highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning,” *Genome Biol.*, vol. 21, no. 1, Jul. 2020, Art. no. 170, doi: 10.1186/s13052-020-02083-3.

[50] T. Tian, J. Zhang, X. Lin, Z. Wei, and H. Hakonarson, “Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data,” *Nature Commun.*, vol. 12, no. 1, Mar. 2021, Art. no. 1873, doi: 10.1038/s41467-021-22008-3.

[51] Z. C. Bin Duan et al., “Learning for single-cell assignment,” *Sci. Adv.*, vol. 6, 2020, Art. no.eab0855.

[52] Y. Lieberman, L. Rokach, and T. Shay, “CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments,” *PLoS One*, vol. 13, no. 10, 2018, Art. no.e0205499, doi: 10.1371/journal.pone.0205499.

[53] Z. Miao, P. Moreno, N. Huang, I. Papatheodorou, A. Brahma, and S. A. Teichmann, “Putative cell type discovery from single-cell gene expression data,” *Nature Methods*, vol. 17, no. 6, pp. 693–703, 2020, doi: 10.1038/s42256-020-00244-4.

[54] K. Boufrea, S. Seth, and N. N. Batada, “scID uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell RNA-seq data with batch effect,” *Science*, vol. 23, no. 3, Mar. 2020, Art. no. 100914, doi: 10.1126/scijs.2020.100914.

[55] L. Wang et al., “An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data,” *Nature Mach. Intell.*, vol. 2, no. 11, pp. 693–703, 2020, doi: 10.1038/s42256-020-00244-4.

[56] X. Guo, W. Li, and F. Iorio, “Convolutional neural networks for steady flow multicellular organism,”*Acta Ophthalmologica*, vol. 357, no. 6352, pp. 661–667, 2017.

[57] J. Hu, X. Li, G. Hu, Y. Lyu, K. Sutskat, and M. Li, “Iterative transfer learning with neural network for clustering and cell type classification in single cell RNA-sequencing analysis,” *Nature Mach. Intell.*, vol. 2, no. 10, pp. 607–618, Oct. 2020, doi: 10.1038/s42256-020-00233-7.

[58] J. Hu, X. Li, G. Hu, Y. Lyu, K. Sutskat, and M. Li, “Iterative transfer learning with neural network for clustering and cell type classification in single cell RNA-sequencing analysis,” *Nature Mach. Intell.*, vol. 2, no. 10, pp. 607–618, Oct. 2020, doi: 10.1038/s42256-020-00233-7.

[59] Y. Madadi, J. Sun, H. Chen, R. Williams, and S. Yousefi, “Detecting retinal neural and stromal cells classes and ganglion cell subtypes based on transcriptome data with deep transfer learning,” *Bioinformatics*, vol. 38, no. 18, pp. 4321–4329, 2022, doi: 10.1093/bioinformatics/btaa514.

[60] Z. Li and H. Feng, “A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data,” *Sci. Rep.*, vol. 12, no. 1, Jan. 2022, Art. no. 910, doi: 10.1038/s41598-021-04473-4.

[61] J. Hu, X. Li, G. Hu, Y. Lyu, K. Sutskat, and M. Li, “Iterative transfer learning with neural network for clustering and cell type classification in single cell RNA-sequencing analysis,” *Nature Mach. Intell.*, vol. 2, no. 10, pp. 607–618, Oct. 2020, doi: 10.1038/s42256-020-00233-7.

[62] Y. Madadi, J. Sun, H. Chen, R. Williams, and S. Yousefi, “Detecting retinal neural and stromal cells classes and ganglion cell subtypes based on transcriptome data with deep transfer learning,” *Bioinformatics*, vol. 38, no. 18, pp. 4321–4329, 2022, doi: 10.1093/bioinformatics/btaa514.

[63] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, “Batch effects in single-cell RNA-sequencing data are corrected by matching nearest neighbors,” *Nature Biotechnol.*, vol. 36, no. 5, pp. 421–427, Jun. 2018, doi: 10.1038/nbt.4091.

[64] C. T. B. Martin Rosvall, “Maps of random walks on complex networks reveal community structure,” *Appl. Math.*, vol. 105, no. 4, pp. 1118–1123, 2006, doi: 10.1037/0003-066x.105.4.1118.

[65] G. X. Zhang et al., “Massively parallel digital transcriptional profiling of single cells,” *Nature Commun.*, vol. 8, Jan. 2017, Art. no. 14049, doi: 10.1038/ncomms14049.

[66] H. L. Pascal Vincent, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep net- work with a local denoising criterion,” *J. Mach. Learn. Res.*, vol. 11, pp. 3361–3408, 2010.

[67] A. Gelman and J. Hill, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge, U. K., MA: USA, Cambridge Univ. Press, 2006.

[68] M. Cortes, J. Bennett, M. B. Brown, and K. A. W. Furtado, “A new method for classifying single cells,” *Science*, vol. 275, no. 5302, pp. 418–420, 1997.

[69] G. Kostopoulos, S. Karlos, S. Kotsiantis, and O. Rago, “Semi-supervised learning for single-cell gene expression analysis,” *Genome Biol.*, vol. 16, no. 1, pp. 1–10, 2015.

[70] J. Cao et al., “Comprehensive single-cell transcriptional profiling of a multicellular organism,” *Science*, vol. 357, no. 6352, pp. 661–667, 2017.

[71] A. Gelman and J. Hill, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge, U. K., MA: USA, Cambridge Univ. Press, 2006.

[72] H. L. Pascal Vincent, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep net- work with a local denoising criterion,” *J. Mach. Learn. Res.*, vol. 11, pp. 3361–3408, 2010.

[73] A. Gelman and J. Hill, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge, U. K., MA: USA, Cambridge Univ. Press, 2006.

[74] M. Cortes, J. Bennett, M. B. Brown, and K. A. W. Furtado, “A new method for classifying single cells,” *Science*, vol. 275, no. 5302, pp. 418–420, 1997.

[75] G. Kostopoulos, S. Karlos, S. Kotsiantis, and O. Rago, “Semi-supervised learning for single-cell gene expression analysis,” *Genome Biol.*, vol. 16, no. 1, pp. 1–10, 2015.

[76] J. Cao et al., “Comprehensive single-cell transcriptional profiling of a multicellular organism,” *Science*, vol. 357, no. 6352, pp. 661–667, 2017.

[77] A. Gelman and J. Hill, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge, U. K., MA: USA, Cambridge Univ. Press, 2006.
Hao Chen received the PhD degree in anatomy from Michigan State University. He is a full professor with the Department of Pharmacology, Addiction Science and Toxicology, University of Tennessee Health Science Center. He has a wide range of research interests, with a focus on genetic and genomics of substance abuse related phenotypes modeled using rats. In addition, his lab contributed to many transcriptome studies of brain regions critically involved in the reward circuitry. Lastly, his lab is working on using deep learning to analyze rat social behaviors.

W. Daniel Stamer received the bachelor of science degree in molecular and cellular biology, and the doctorate degree in pharmacology and toxicology from the University of Arizona, in 1990 and 1996, respectively. After completing two research fellowships, he started his research program in 1998 with the University of Arizona, where he remained for 13 years; rising through the ranks to full professor and director of Ophthalmic Research. He was subsequently recruited to Duke University in 2011, where he currently serves as the Joseph A.C. Wadsworth professor of ophthalmology and professor of biomedical engineering. The primary research focus of the Stamer Laboratory is to understand the molecular and cellular mechanisms that regulate conventional outflow such that novel targets can be identified, validated and used for the development of therapeutics that target/modify the diseased tissue responsible for elevated intraocular pressure in glaucoma. Over the past 30+ years, he has pioneered the development of cellular, tissue, organ culture and murine model systems for use by his laboratory and others to study conventional outflow physiology and pharmacology. His laboratory has worked closely with industry, assisting in the development/pre-clinical testing of several new classes of glaucoma drugs that target the diseased conventional outflow pathway responsible for ocular hypertension. Research progress is documented in more than 175 peer-reviewed primary contributions to the literature and two dozen reviews/book chapters/editorials, having more than 9,900 citations. His work was recognized by the Rudin Prize for Glaucoma in 2012 and the Research to Prevent Blindness Foundation in 2013. Recently, he was elected as ARVO trustee in 2015 and ARVO president in 2019/20 and elected into the Glaucoma Research Society in 2022. He currently holds prominent editorial positions in three premier ophthalmology journals: the Journal of Ocular Pharmacology and Therapeutics, Investigative Ophthalmology and Visual Science and Experimental Eye Research. Moreover, he currently serves on the scientific advisory boards for six companies and three foundations that support glaucoma research.

Robert W. Williams received the BA degree in neuroscience from the University of California, Santa Cruz, in 1975, and the PhD degree in system physiology from the University of California, Davis, in 1983. He did postdoctoral work in developmental neurobiology with Yale where he developed novel stereological methods to estimate neuron populations. In 1989, he moved to the University of Tennessee and in 2013 established the Department of Genetics, Genomics and Informatics, College of Medicine. He holds UT Oak Ridge National Laboratory governor’s chair in computational genomics. He is director of the Complex Trait Community (www.complextrait.org) and founding editor-in-chief of the Frontiers in Neurogenomics.

Siamak Yousefi (Senior Member, IEEE) received the PhD degree in electrical engineering from the University of Texas at Dallas, in 2012, and completed two postdoctoral trainings with the University of California Los Angeles (UCLA) and the University of California San Diego (UCSD). He is an Assistant Professor with the Department of Ophthalmology and the Department of Genetics, Genomics, and Informatics of the University of Tennessee Health Science Center (UTHSC) in Memphis. He is the director of the Data Mining and Machine Learning (DM2L) Laboratory, UTHSC working on broad applications of Artificial Intelligence (AI) in vision and ophthalmology particularly glaucoma diagnosis, prognosis, and monitoring. He is a member of the ARVO.