Integrated Ocular Million Single Cell Meta-Atlas for Discovery and Development

Vinay Swamy1, Zachary Batz2, Temesgen Fufa3, Robert Hufnagel3, and David McGaughey1,✉

August 03, 2020

PURPOSE: To curate a accessible and reliable ocular cell-type specific transcriptome dataset of over one million normal vertebrate ocular cell types with a matching high performance reactive web application to query gene expression across cell type, species, study, and other factors. METHODS: We queried several public repositories to find eighteen healthy, non-perturbed, ocular tissue single cell RNA-seq studies As a non-ocular reference, we also used the pan-mouse Tabula Muris resource. The XX ocular and XX Tabula Muris cells were sent into a Snakemake-based reproducible pipeline we wrote to quantify transcripts, remove cells with poor quality, normalize expression values, create a batch-corrected low dimensional space, performs XX differential expression tests, and outputs all as a single SQLite database file: the Single Cell Eye in a Disk (scEiaD) dataset. XXX thousand published cell labels assignments were curated and a machine learning projected onto the non-labelled cells. Furthermore, we wrote the web application PLAE (<https://plae.nei.nih.gov>) to display scEiaD. RESULTS: The new portal provides quick visualization of vertebrate eye-related transcriptomes published to date by gene/transcript, XX cell types, XX ocular tissues, XX body tissues. As a test of the value of this unified pan-eye dataset, we showed XX. CONCLUSION: The PLAE v1.0 web app serves the pan-ocular and body dataset, scEiaD. This offers the eye community a powerful and quick means to test hypotheses on human gene and transcript expression across 54 body and 19 eye tissues. Futhermore the XXX labelled cells across XX studies, X techologies, and 3 species provides a highly challenging benchmarking resource for single cell algorithm development.

1 Bioinformatics Group, Ophthalmic Genetics & Visual Function Branch, National Eye Institute, National Institutes of Health  
2 Neurobiology, Neurodegeneration & Repair Laboratory, National Eye Institute, National Institutes of Health  
3 Medical Genetics and Ophthalmic Genomics Unit, National Eye Institute, National Institutes of Health

✉ Correspondence: [David McGaughey <[mcgaugheyd@mail.nih.gov](mailto:mcgaugheyd@mail.nih.gov)>](mailto:mcgaugheyd@mail.nih.gov)

# Introduction

## Many published ocular atlases

## Several have accompanying apps

## Apps are generally slow and have minimal exploration tooling

## If validating / exploring a gene expression pattern, crucial to have reproducibility across disparate groups to enhance confidence that effect is biological

## Single cell datasets getting larger

Valentine Svensson figure? MOCA atlas (mouse 2e6 cells?)

## Thus crucial for algorithms to be highly performant

Our multidimensional atlas of 1e6 cells will be a useful reference dataset in: - batch effect correction tools - clustering algorithms - cell type prediction / projection - trajectory inference

## What we did

We have built an integrated meta-atlas that:

Has a batch corrected lower dimensional space (8 dimensions) from scVI

Curated pre-labelled cell types for >400,000 cells.

Projected these onto the unlabelled cells.

Made 77 clusters - subclustered each to create XX total clusters

Pre-calculated pseudobulk differential expression testing in 9 different testing schemes:

## Will do???

Trajectory

Velocity

# Methods

## Identify ocular scRNA studies

SRA EBI Pubmed Found 18 (?) studies

## pan mouse as non-ocular reference

Tabula Muris

## Quantification

kallisto/bustools

## Batch normalization

Tested: - scVI - CCA - insct - combat - nothing - magic - scanorama - harmony - fastMNN

## Grid search to pick optimal parameters

Varied: - number of Hyper Variable Genes (HVG): 1000, 2000,5000,1000 - dims: 4,6,8,10,20,30,50,100 - normalization: scran, seurat standard, libSize, sqrt, counts (scVI only) - nneighbors (UMAP viz only) - knn: 5,7,10 (clustering)

benchmarked with: - LISI - ARI - Silhouette - PCA variance explained before/after correction

## Clustering

Louvain-jaccard from Seurat

## 

# Results

## pseudobulk differential expression testing

* celltype (published) against remaining
* celltype (published) pairwise against celltype (published)
* species specific within celltype
* same for celltype (projected)
* same for cluster

# Conclusion