Introduction

The chosen paper, "scRNA-seq uncovers the transcriptional dynamics of *Encephalitozoon intestinalis* parasites in human macrophages" from Jaroenlak et al. (June 2024), investigates how this specific parasite manipulates macrophages to promote its survival and cause infection. *Encephalitozoon intestinalis* is a microsporidia, a single-celled parasite that is able to replicate within host cells, and is one of the most common microsporidia that infect humans. Transmission typically occurs through contaminated food or water, meaning initial infection occurs within the small intestine, and can be fatal to immunocompromised patients.

Once in the small intestine, *E. intestinalis* invades the lining epithelial cells before spreading to multiple cell types, including macrophages. A couple methods employed by *Encephalitozoon* to evade their destruction include inhibiting the macrophages' processes of phagosome acidification, fusion with lysosomes, and apoptosis, all while promoting their own replication within the host macrophage. This mechanism is not entirely understood and was studied by the authors through single cell RNA sequencing of macrophages infected with *E. intestinalis*, with both host and parasite transcriptomes being analyzed. Previous studies have employed bulk RNA sequencing to study these transcriptional dynamics, though these analyses only provide an overview of the transcriptomics throughout the entire cell population. Single cell RNA sequencing is representative of the complexity and heterogeneity that is associated with infection in specific cells, and will not be biased by low infection rates or out of sync replication.

Methods

The main steps of the code are emphasized in headings before the code blocks: Read in data, Demultiplex, remove doublets & negatives, and separate donors 3 & 4, QC, Common PCA Prep for Both Figures, Run PCA for Downstream Visualization, Plotting for Figure 2 & 5, then EnrichR Pathway Analysis. Reading in the data included processing tsv and mtx files that were downloaded from the GEO with accession number GSE268707. There were 9 files total (3 for each: donor 1, donor 2, and combined donor 3 & 4) and the function implemented pandas and scipy.

The next section encasing demultiplexing was done based on the hashtag antibodies, marking the infection timepoint the cells were associated with. The hashsolo function from scanpy.external was then able to classify the cells as a "Singlet", "Doublet", or "Negative", in which "Doublet" and "Negative" were filtered out. This function also handled labeling the cells in a new column with their donor sample. This would help for the subsequent separation of the donor 3 and donor 4 combined dataset. Quality control was next performed by removing the cells that did not pass the set thresholds. This was implemented in python using scanpy's calculate_qc_metrics. Each donor had different thresholds, as specified in the paper and Jupyter Notebook, making it necessary to perform the donor 3 and donor 4 separation prior. Additional columns of 'species', to identify human transcripts from parasites, and 'pct_parasite', to determine infected cells from uninfected, were also added at this stage to assist with subsetting. The 4 AnnData objects post-QC were used to create an additional list, this time of the QC data subsetted by 'human' in 'species'.

The result is two lists, each of 4 AnnData objects. These two lists moved onto the PCA prep which included log normalization, analyzing each AnnData for highly variable genes (HVGs) via scanpy, then integrating (batch-effect correcting) all 4 into one AnnData object using scanorama.correct_scanpy. This final AnnData object was then scaled by a Z-score transformation with scanpy's scale function. These two results, one for figure 2 and one for figure 5, were analyzed with a common scanpy workflow involving PCA, finding nearest neighbors, UMAP, and leiden clustering within the Run PCA section. Here, the paper's parameters were matched for figure 2 data; 41 PCs were used in nearest neighbor

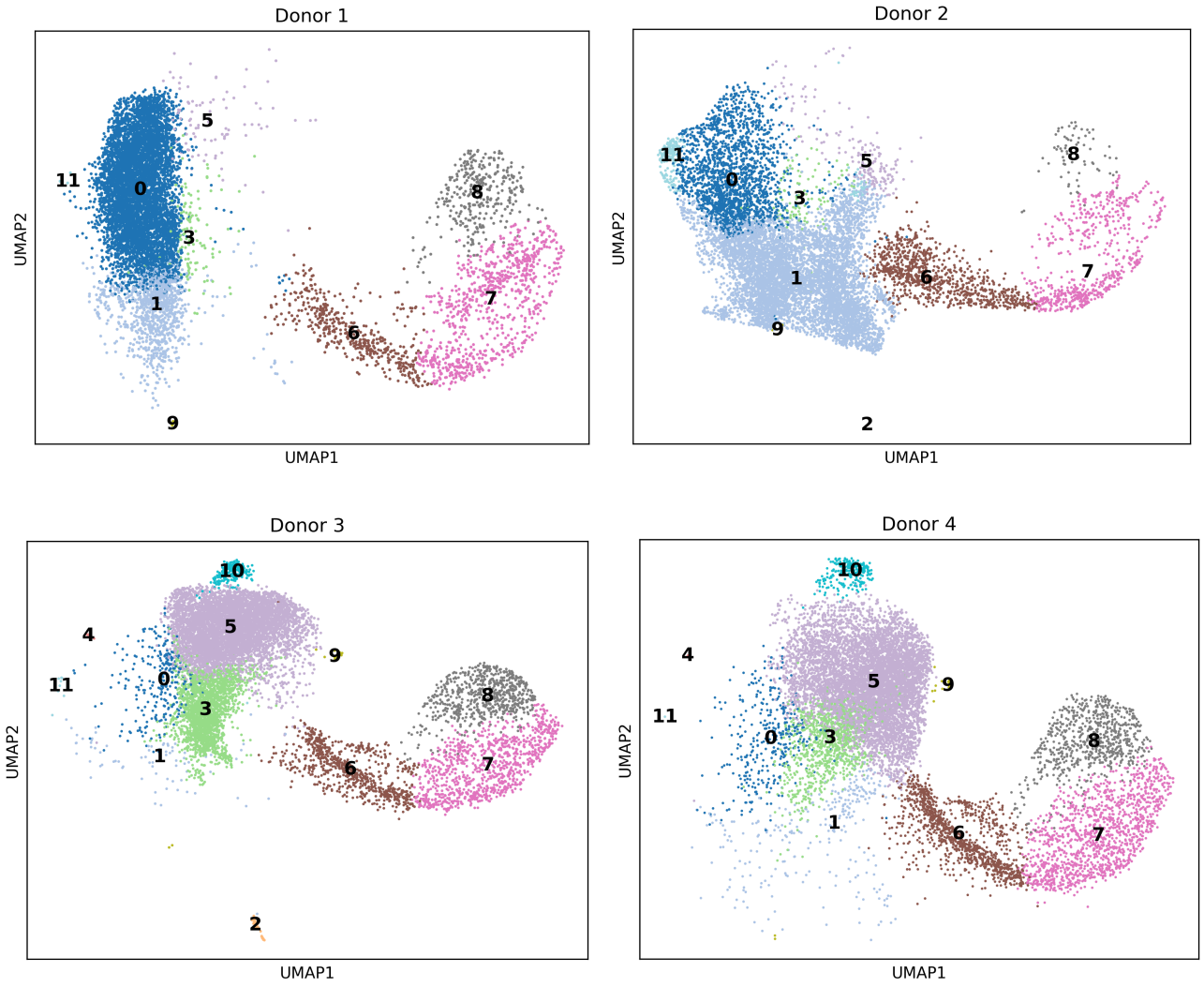
calculations, and clustering was at a resolution of 0.4 to produce 12 clusters. For figure 5, 43 PCs were used as in the paper, though their resolution of 0.2 resulted in only 6 clusters for this workflow. To achieve 9 clusters, the resolution was increased to 0.43.

For visualization, the final dataset for figure 2 was subset into 4 different datasets based on their 'donor' column, making 4 separate UMAP plots. The dataset for figure 5 was subset based on 'pct_parasite', to isolate uninfected cells (those with < 2% parasite transcripts) from all the cells, creating 2 separate UMAP plots. The two plots from figure 5 showed the largest difference occurred in cluster 7. These genes from this cluster were isolated and analyzed in EnrichR. The paper had conducted a differential gene expression analysis, and filtered by a fold change > 1.5, but the integration proved to cause problems with the scanpy.tl.rank_genes_groups function. As a work around, pathway analysis was instead run on all genes from cluster 7, and the top 4 pathways were plotted by their -log10(p-value).

Results

The first figure recreated was figure 2b-e, showing 4 UMAP plots of the full transcriptome, host and parasite, cells separated by donor. The recreation within the scanpy workflow is shown below in Figure 2b-e Recreation.

Figure 2b-e Recreation



As the paper saw, these UMAPs cluster into 12 different clusters, and create two different larger cell populations that the paper termed population A (left) and population B (right). There are three clusters

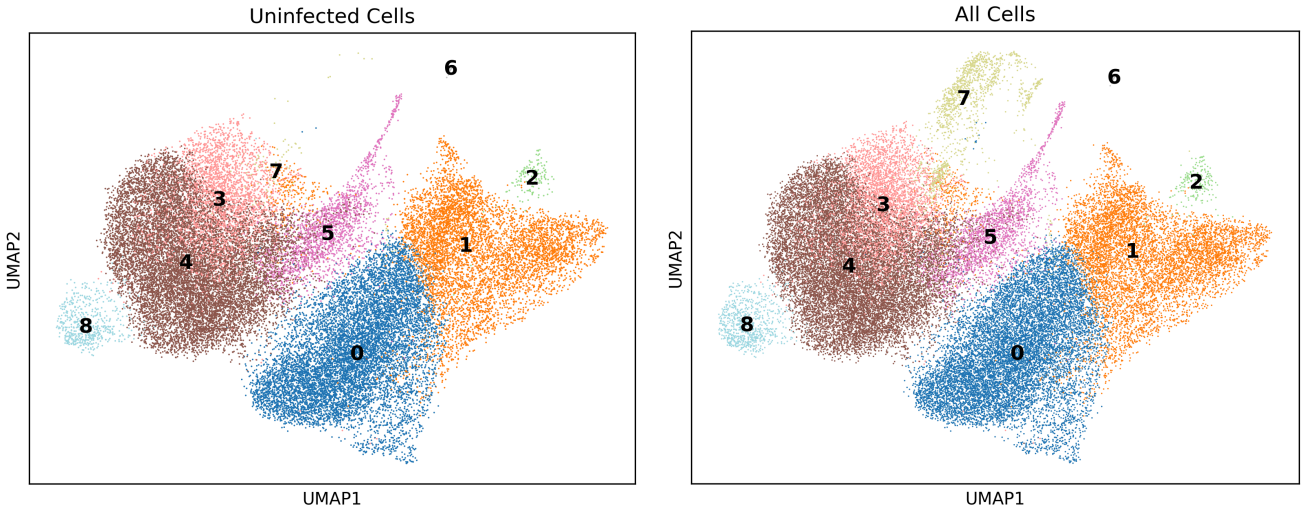
that make up population B in both the paper and this recreation, with the rest composing population A. The paper shows population A is made of both infected and uninfected cells, while population B is made up of infected cells. These results support this general idea as population A is consistently more diverse than population B. Not all clusters are accounted for within every donor, however.

It is apparent how different the UMAPs look across all four donors. As the paper used Seurat in R for the analysis, the work done here was in Python's Scanpy package. Though there are analogous functions and similar workflows, the underlying algorithmic differences and computations come through within the data. For example, from the initial reading in of the datasets, the same amount of cells were included as the paper's supplementary table lists for each donor, confirming the correct data was downloaded. This quickly changed as after QC and demultiplexing, each donor set was left with 400-600 more cells than the paper had after their own QC. Despite making the thresholds more stringent at various values and setting specific parameters for the demultiplexing, there was little to no change in the amount of cells that passed the current QC.

Another observation is donors 1 and 2 are more comparable to one another, just as donors 3 and 4 are to each other; donors 1 and 2 lack the clusters that donors 3 and 4 primarily have, and vice versa. This could potentially be because donors 3 and 4 were in a combined dataset during demultiplexing, while donors 1 and 2 were separate. If donors 3 and 4 were separate from the beginning, they could have appeared to be more similar to donors 1 and 2.

To achieve reproducing the pathway analysis bar chart visual in figure 5, two UMAPs first had to be created to compare the clustering between uninfected hosts and infected hosts. These are depicted in Figure 5a Recreation.

Figure 5a Recreation

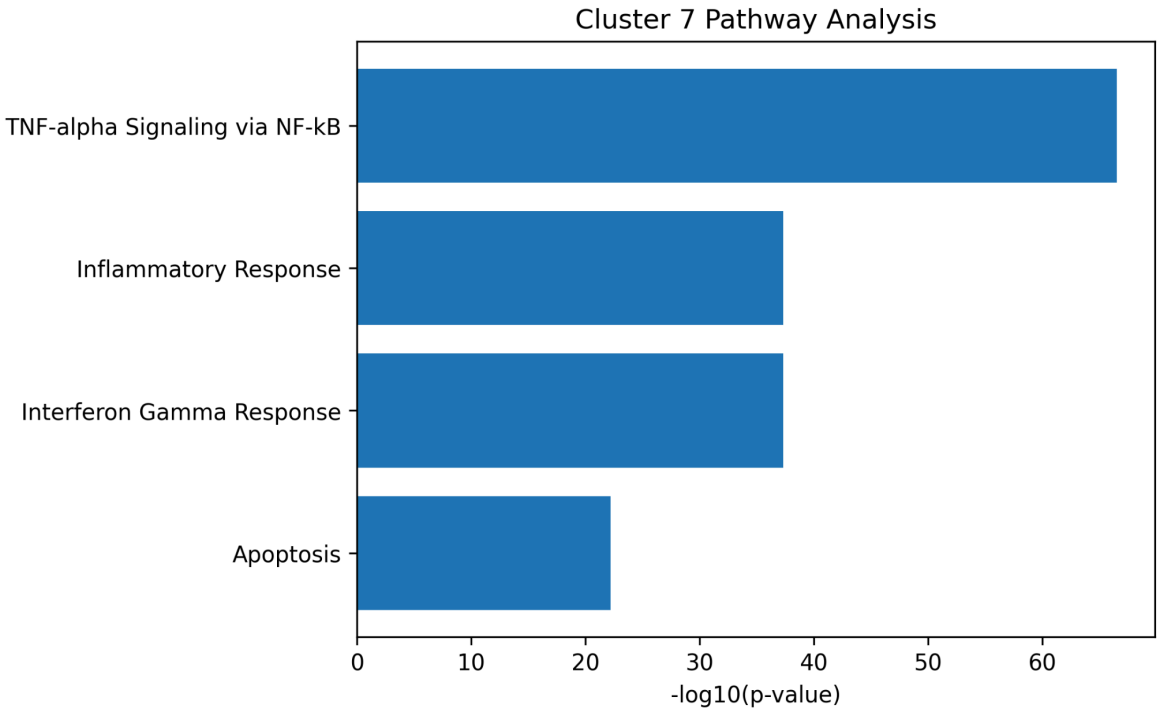


These UMAPs are very similar to one another with each cluster 0-8 appearing in the same area, and all clusters being represented within both maps. The main difference seen here is the size of cluster 7, which is nearly nonexistent in the uninfected cell subset, but is a defined cluster within the mixed cell types, implying its identity of infected cells. This cluster was then specifically isolated for the pathway analysis.

One limitation of analysis regarding cluster 7 was being unable to replicate the differential gene expression analysis prior to EnrichR pathway analysis. An important log normalized data layer within the AnnData objects was unable to be preserved through highly variable gene splicing and integration

despite multiple attempts throughout different functions within the pipeline. Due to not having differentially expressed genes, all 583 genes from cluster 7 were exported for EnrichR analysis online. Results were then exported and graphed, displayed in Figure 5g/h Recreation.

Figure 5g/h Recreation



A few of the genes that contributed to this result included LY6E, TREM2, and gene families of CCL and CD, all involved in immune regulation, recognition, and signaling.The genes the authors chose to emphasize from their data, INHBA, IER3, CIR1, MALAT1, and NEAT1, are present within the gene list as well.

These pathways cannot be determined to be upregulated or downregulated, but the paper found similar pathways when they filtered for upregulated genes. Specifically, TNF-alpha Signaling via NF-kB, inflammatory response, and interferon gamma response were found to be upregulated in both clusters analyzed by the authors. It is likely the genes found in cluster 7 were primarily upregulated targets in infected cells. The fourth pathway found, apoptosis, is interesting as macrophages infected with *E. intestinalis* have been shown to inhibit the host cells' ability to undergo apoptosis. This result could therefore be from the remaining genes that are found to be downregulated during infection.

The described pipeline aimed to replicate a single cell RNA analysis from Seurat within Scanpy using sequencing data from human macrophages infected with *E. intestinalis*. The differences between the two methods were apparent from the start, as scanpy is seemingly more lenient with its filtering/demultiplexing than Seurat is. This set up the stark differences that were to follow when it came to the visualizations, though the underlying biology proved to be more robust.