Analysis of Gene Expression Patterns in Multiple Sclerosis

Master Degree in Computer Science, University of Pisa
Computation Healt Laboratory course (755AA), Prof. Corrado Priami
Academic Year: 2023-2024

Gabriele Benanti
g.benanti@studenti.unipi.it
Roll number: 550552

Giulia Ghisolfi
g.ghisolfi@studenti.unipi.it
Roll number: 664222

Alessandro Stefanelli
a.stefanelli3@studenti.unipi.it
Roll number: 686084

Abstract

In this report, we present our methodology and the results we obtained on the differential expression analysis in multiple sclerosis (MS) patients compared to healthy controls across data collected from different studies and datasets. We investigate these expression differences in peripheral blood mononuclear cells (PBMCs) and cerebrospinal fluid cells (CSFs), focusing our analysis on various cell types, particularly T cells, B cells, Monocytes, Plasma cells, and Innate Lymphoid Cells (ILC). Our findings reveal significant differences in gene expression profiles, with a notably higher number of downregulated genes in MS patients compared to healthy controls, especially within the CSF, highlighting the complex alterations in cellular functions associated with MS pathology.

1 Introduction

Multiple sclerosis is a chronic inflammatory, neurodegenerative, and autoimmune disease that primarily affects young adults, more frequently women [2] [9]. Despite extensive research, the underlying cause and mechanisms behind multiple sclerosis remain elusive, though complex gene–environment interactions are believed to play a significant role. The current understanding of multiple sclerosis pathogenesis involves multilayered interactions between genetic and environmental factors [1].

Despite these advances, studying multiple sclerosis remains challenging due to the lack of comprehensive data. Most published studies have small sample sizes that are not statistically representative, often consisting of patients from a single hospital who receive similar treatments and share demographic characteristics such as age and gender.

This difficult in obtaining a significant sample is partly because multiple sclerosis is a heterogeneous disease with a variable and unpredictable course. Due to its complex nature, multiple sclerosis is difficult to diagnose, and responses to specific treatments can vary significantly between individuals [5].

In our project, we aimed to address these challenges by integrating data from different datasets to identify biomarkers, genes related to specific cell types, in patients with multiple sclerosis via differential expression analysis. One of the primary obstacles we encountered was the scarcity of a usable dataset, which necessitated extensive data acquisition efforts.

This report details our methodology, the obstacles we faced, and the insights gained from our integrative approach to biomarker identification in multiple sclerosis patients.

2 Datasets

In this section, we provide an overview of the datasets selected and utilized in our project. We analyzed various studies and publicly available datasets, focusing on their origins, data availability, types of data provided, previous studies in which they were used, exploratory data analysis (EDA) conducted on each dataset, and the preprocessing steps we undertook.

A summary of the number of cells we have available and the number of patients from whom they were extracted from the first three datasets containing transcriptomic data is reported in Table 1.

	MS			HC		
	patients	**PBMC**	**CSF**	**patients**	**PBMC**	**CSF**
GSE239626	10	72,317	-	-	-	-
GSE194078	3	17,083	17,133	9	44,398	31,277
GSE138266	6	25,553	19,306	3	15,604	5,641
Total	19	114,953	36,389	12	60,002	36,918

Table 1: Summary of the number of cells we have available and the number of patients from whom they were extracted. PBMC refers to Peripheral Blood Mononuclear Cells, and CSF refers to Cerebrospinal Fluid. MS stands for Multiple Sclerosis, and HC stands for Healthy Control.

2.1 GSE239626

The first dataset we utilized, GSE239626[1], contains data exclusively from multiple sclerosis patients and was the only one available to us that included antibody data. This dataset provided a crucial starting point for our analysis, allowing us to extract cell type labels for each cell, since, none of the datasets, including this one, initially provided cell type labels.

These labels, derived from the antibodies counts associated with respective cell types, were then utilized to infer cell type labels in other datasets.

The dataset was collected to study the benefits of vitamin D for patient-important outcomes among people with MS, as detailed in [4] and [6]. These studies involved administering vitamin D supplementation to some patients, in addition to their existing treatments, while others received a placebo. However, none of the studies reported positive outcomes. Additionally, the other therapies received by the patients were not mentioned, making it impossible for us to categorize patients based on their treatments.

Therefore, this dataset was used solely as a collection of data from MS patients, which was later compared with data from healthy controls in other datasets, without considering the specific treatments each patient underwent, due to the lack of comprehensive information.

A comprehensive description of the publicly available data on GEO (Gene Expression Omnibus) related to this dataset and the studies conducted on them is provided in Table 6 in Appendix.

[1]The dataset is accessible on GEO at the following link: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE239626`

Exploratory Data Analysis

The dataset comprises data from 10 distinct patients, each of whom provided two samples taken three months apart. These samples were processed to extract peripheral blood mononuclear cells (PBMCs), a type of lymphocyte. Following a protocol for live lymphocytes, gene expression (GEX) libraries and cell-surface protein expression libraries were constructed.

The dataset comprises a total of 72,317 cells, divided into 20 samples, two for each of the 10 patients, for which counts of 36,601 genes and 35 antibodies are reported.

In Table 10 in the Appendix, summary statistics of total gene counts per cell for each of the 10 patients in the dataset are provided. The table shows that the number of cells per sample ranges from 2956 to 5491. There is some variability in the average number of cells per patient, indicating potential differences in sample quality or biological variability among patients.

The violin plots in Figure 1 below illustrate the distribution of total gene counts detected per cell for each patient in the dataset. From the plots, it is evident that the means vary across individuals. Particularly, for the first and sixth patients, the means are higher compared to the rest of the samples. For the other individuals, a similar distribution, although slight variations can be observed in the quantiles for each of them, and the tails of the violin plots also show variability.

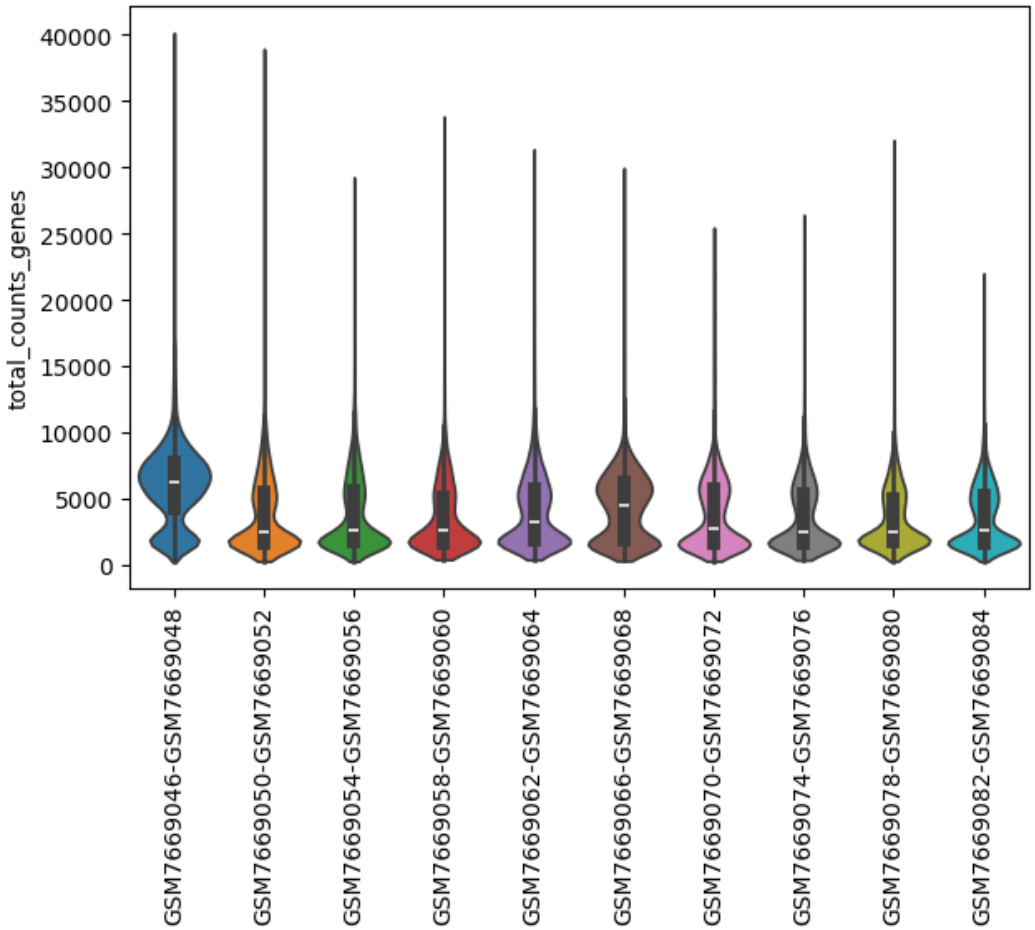


Figure 1: Violin plots illustrate the distribution of the total number of genes detected per cell for each patient in the GSE239626 dataset.

In Figure 2 the plots of the distributions of detected genes and antibody per cell across the dataset are shown.

Examining the distribution of the total number of detected genes per cell across the dataset, we observe that the majority of cells have fewer than 5000 genes detected. As the distribution progresses, there are progressively fewer cells with over 20,000 total detected genes, reaching a maximum of 40,000.

A similar analysis was conducted for the antibodies, where we also observed higher peaks for lower count values, typically below 20,000. There were only a few cells with counts surpassing 50,000 (with a maximum exceeding 250,000). Additionally, the distributions exhibit varying degrees of skewness towards their tails, with some exhibiting longer tails than others.

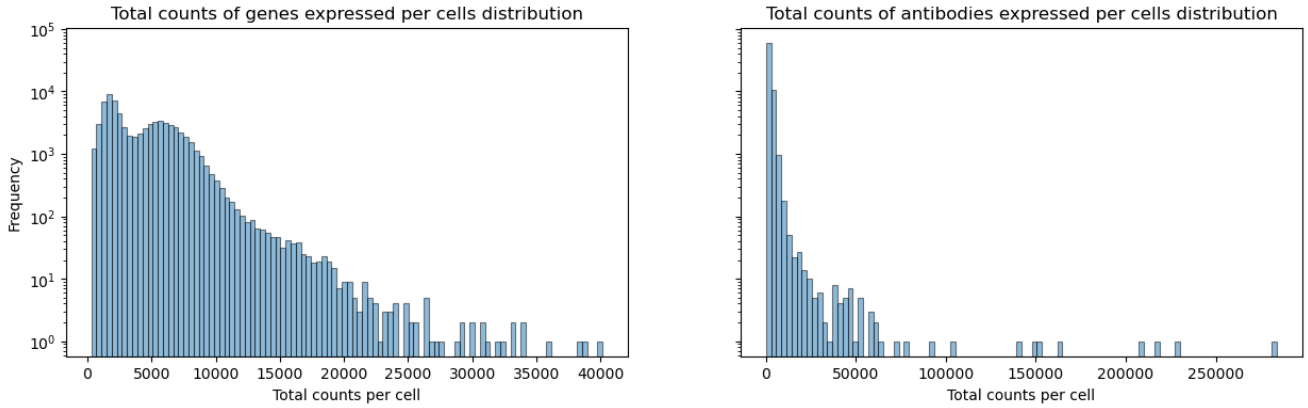


Figure 2: Distribution plots of total detected gene (left) and antibody (right) per cell across all samples.

We also studied the genes and proteins with the highest average expression levels across all cells in the provided dataset, providing insights into the dominant expression patterns. The gene with the highest expression is MALAT1, followed by mitochondrial genes such as MT-CO3, MT-CO2, MT-ATP6, and others. Other notable genes include B2M, RPS27, EEF1A1, and several ribosomal genes like RPL41, RPL10, RPL13, RPS12, and RPLP1, as seen in Figure 3.

Similarly, we examined the proteins with the highest expression across all cells in the dataset, as shown in Figure 4. The protein with the highest expression is CD11a, followed by CD3, CD8, CD45RA, and CD4. Other notable proteins include CD27, CD20, IgM, HLA-DR, as well as CD49d, CD62L, CD24, and CD38.

2.2 GSE194078

GSE194078[2] is a transcriptomic dataset containing only gene expression data.

A comprehensive description of the available data on GEO related to this dataset and the studies conducted on them, which was explored in detail in [7], is provided in Table 7 in Appendix.

Exploratory Data Analysis

The publicly available dataset contained samples from 12 distinct patients, including 3 MS patients and 9 non-MS patients used as healthy controls. In total, the dataset comprised 109,891 cells, with 75,675 cells from healthy control samples and 34,216 cells from MS samples.

In Figures 5 and 6, violin plots illustrate the distribution of gene counts in PBMC and CSF cells, respectively, from the GSE194078 dataset. Distributions corresponding to patients `GSM5827377`,

[2]The dataset is accessible on GEO at the following link: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194078`

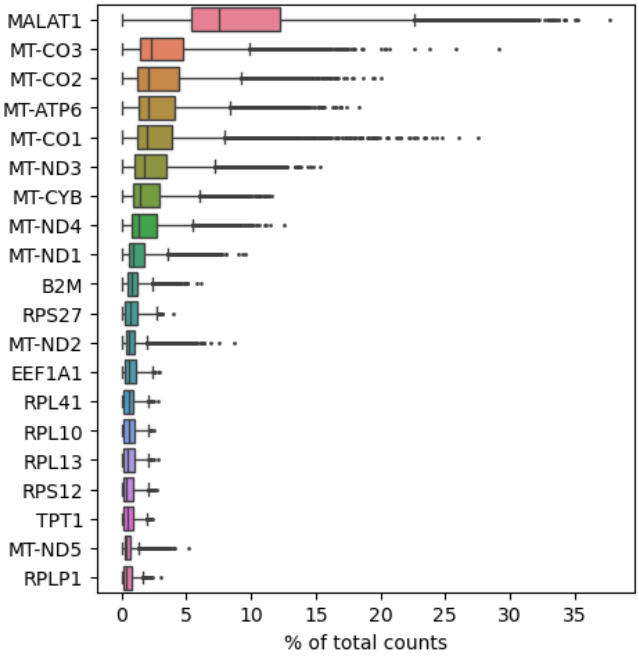


Figure 3: Box plot illustrating the distribution of genes expression levels for the top 20 genes in dataset GSE239626 with the highest average expression across all cells in the dataset. The gene with the highest expression is MALAT1, followed by mitochondrial genes such as MT-CO3, MT-CO2, MT-ATP6, and others. Other notable genes include B2M, RPS27, EEF1A1, and several ribosomal genes like RPL41, RPL10, RPL13, RPS12, and RPLP1.

`GSM5827379`, and `GSM5827385` represent MS patients, while the remaining distributions correspond to healthy controls.

From the distributions shown in the violin plots, we observe differences in the gene count distributions for each individual. However, it is difficult to discern a clear difference between the distributions of MS patients and healthy controls for both PBMC and CSF cells. This suggests that while there are individual variations, a clear distinction based solely on gene count distributions between MS patients and healthy controls is not evident from these plots.

Figures 7 represent a box plot illustrating the distribution of gene expression levels for the top 20 genes with the highest average expression across all cells in the dataset. This box plot effectively visualizes the distribution of total counts among the most highly expressed genes, with MALAT1 standing out as the most predominant. The presence of mitochondrial and ribosomal genes underscores their essential roles in cellular function and metabolism.

The remaining genes, including B2M, RPS27, MT-CO1, MT-ATP6, EEF1A1, RPL10, RPL41, RPL13, MT-CO3, MT-CO2, TPT1, TMSB4X, RPS12, RPLP1, MT-ND3, RPS18, MT-CYB, RPS29, and RPL30, exhibit much lower percentages of total counts, all staying below 10%. Many of the genes listed (e.g., MT-CO1, MT-ATP6, MT-CO3, MT-CO2, MT-ND3, MT-CYB) are mitochondrial genes, indicating a significant contribution of mitochondrial activity to total gene expression. Genes like RPS27, RPL10, RPL41, RPL13, RPS12, RPLP1, RPS18, RPS29, and RPL30 are ribosomal proteins, highlighting the importance of protein synthesis processes.

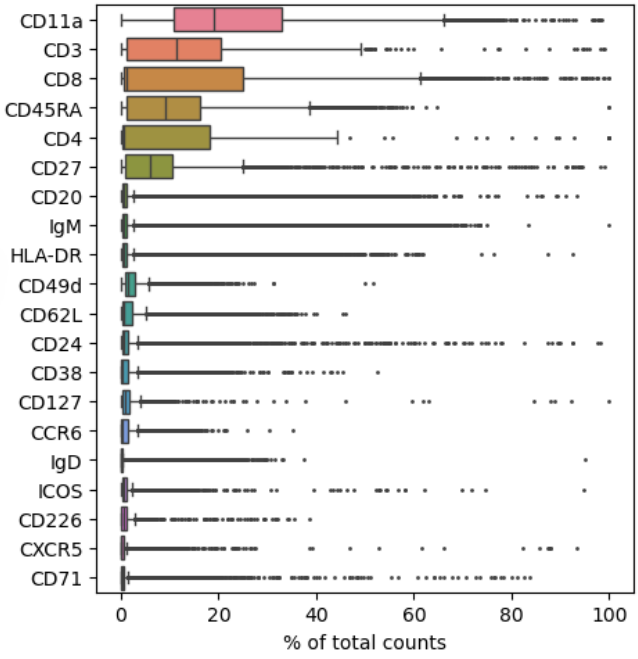


Figure 4: Box plot illustrating the distribution of antibodies expression levels for the top 20 proteins in dataset GSE239626 with the highest average expression across all cells in the dataset. The antibody with the highest expression is CD11a, followed by CD3, CD8, CD45RA, and CD4. Other notable proteins include CD27, CD20, IgM, HLA-DR, as well as CD49d, CD62L, CD24, and CD38.

2.3 GSE138266

The third dataset with transcriptomic data with gene counts that we considered is the GSE138266[3], which includes both cerebrospinal fluid (CSFs) and peripheral blood mononuclear cells (PBMCs) of multiple sclerosis (MS) patients compared to healthy controls (HC).

This dataset was analyzed in a study published in [12]. A summary of the dataset and the study can be found in Table 8 in Appendix.

Exploratory Data Analysis

The dataset available on GEO consists of samples from 6 healthy control (HC) individuals and 6 multiple sclerosis (MS) patients for cerebrospinal fluid (CSF) analysis. Additionally, it includes samples from 5 HC individuals and 5 MS patients, corresponding to those for whom CSF samples were reported, for peripheral blood mononuclear cells (PBMC) analysis.

The dataset encompasses a total of 814,177 cells.

Violin plots depicting the distribution of the total number of detected genes per CSF and PBMC cell for each patient in the dataset are presented in Figures 8 and 9.

Considering the extensive number of cells available, managing the dataset became impractical and computationally challenging. After evaluating the distributions, we opted not to utilize the data pertaining to CSF cells from patients PST.

In total, we utilized 25,553 PBMC cells and 19,306 CSF cells from MS samples. Additionally, for healthy controls (HC), there are 15,604 PBMC cells along with 5,641 CSF cells.

[3]The dataset is accessible on GEO at the following link: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138266`

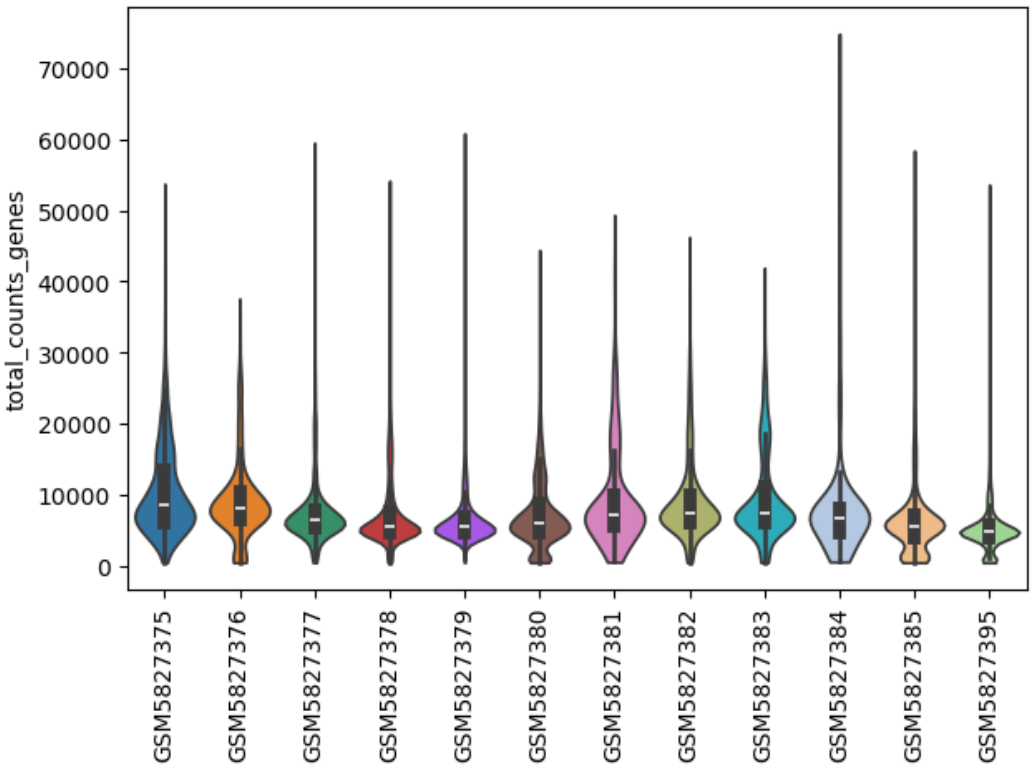


Figure 5: Violin plots illustrate the distribution of the total number of genes detected per PBMC cell for each patient in the GSE194078 dataset. Distributions corresponding to patients `GSM5827377`, `GSM5827379`, and `GSM5827385` represent MS patients, while the remaining distributions correspond to healthy controls.

2.4 GSE173787

To complement our gene expression analysis conducted on the transcriptomic dataset previously described, we integrated our findings with data from the GSE173787[4] dataset.

For a comprehensive understanding of the study conducted with this dataset, refer to [9]. Furthermore, in Table 9 in the Appendix, essential information about the study and the dataset is provided.

Among the processed and published data, there is a dataset that identifies Differentially Expressed Genes (DEGs) through transcriptomic analysis of CD19+ B cells at the onset of disease. This dataset sheds light on alterations in the complement activation pathway and pathways related to B lymphocyte differentiation and activation. The samples were obtained from PBMC cells.

In the processed dataset used, there are 547 genes differentially expressed between Multiple Sclerosis (MS) patients and Healthy Controls (HC). For each gene, the following values are provided to characterize its expression pattern: logFC (Log Fold Change), logCPM (Log Counts Per Million), p-value and FDR (False Discovery Rate).

A summary of statistics of logFC, logCPM, p-value, and FDR for the 547 differentially expressed genes is reported in Table 2.

[4]The dataset is accessible on GEO at the following link: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE173787`

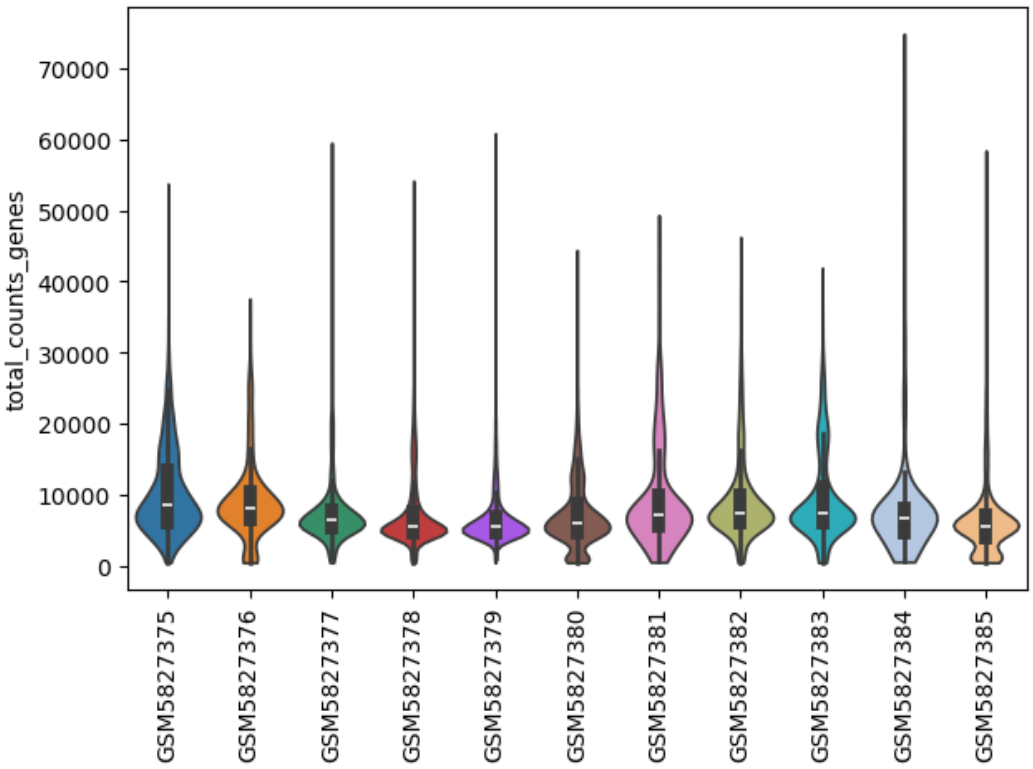


Figure 6: Violin plots illustrate the distribution of the total number of genes detected per CSF cell for each patient in the GSE194078 dataset. Distributions corresponding to patients `GSM5827377`, `GSM5827379`, and `GSM5827385` represent MS patients, while the remaining distributions correspond to healthy controls.

	logFC	**logCPM**	**P value**	**FDR**
mean	0.8677	3.6941	6.74×10^{-4}	1.53×10^{-2}
std	1.1687	2.2032	8.42×10^{-4}	1.38×10^{-2}
min	-3.1006	-0.3859	6.06×10^{-13}	1.01×10^{-8}
25%	0.5944	1.8440	3.57×10^{-5}	3.32×10^{-3}
50%	1.0172	3.5614	2.64×10^{-4}	1.07×10^{-2}
75%	1.6418	5.1479	1.07×10^{-3}	2.52×10^{-2}
max	6.6208	12.0595	3.26×10^{-3}	4.99×10^{-2}

Table 2: Summary statistics of logFC (log fold change), logCPM (log counts per million), p-value, and FDR (false discovery rate) for the 547 differentially expressed genes in dataset GSE173787.

3 Cell Type Extraction from Antibody Clustering

3.1 Antibody Clustering

For this task, we utilized the ScanPy[5] library. We employed all 35 antibodies available in the dataset GSE239626 along with the total number of cells at our disposal (72,317 samples collected from 5 distinct patients).

The choice to utilize all the 35 features is due to both the relatively small number of them and the complementary information they provide. The correlation matrix among antibody features is

[5] `https://scanpy.readthedocs.io/en/stable/`

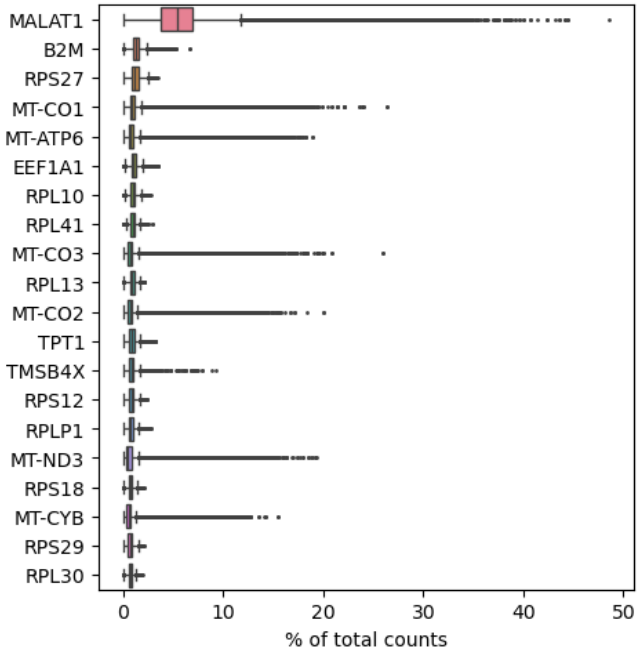


Figure 7: Box plot illustrating the distribution of gene expression levels for the top 20 genes in dataset GSE194078 with the highest average expression across all cells in the dataset. This plot effectively visualizes the distribution of total counts among the most highly expressed genes, with MALAT1 standing out as the most predominant. The remaining genes exhibit much lower percentages of the average fraction of counts assigned to each gene across all cells, all staying below 10%. This distribution showcases the dominance of one highly expressed gene while the majority maintain lower expression levels.

shown in Figure 10. While there are areas of high correlation (both positive and negative), the presence of many areas with low or no correlation indicates that these features are not entirely redundant. By including all features, we can improve the robustness of predictive models by considering the full spectrum of antibody features. The correlation matrix demonstrates that while some features are correlated, there is enough diversity and unique information among the features to justify their inclusion.

We also considered the logarithmic variance of the antibodies. Figure 11 shows the plot of the logarithmic variance of the antibody capture features. Despite some features having lower variances, it is essential to consider all of them initially to ensure a comprehensive overview of potential biological differences. Features with low variance might still provide complementary information when used alongside others.

Initially, we reduced the dimensionality of the data using principal component analysis (PCA) with ARPACK as the singular value decomposition solver, which allows for partial singular value decomposition of a sparse matrix. We set the number of principal components to compute to 30.

We opted to utilize the 5 principal components for clustering, as further increasing the number of dimensions did not significantly improve performance but increased the computational time required for the algorithm execution.

The clustering algorithm chosen and evaluated as optimal from our preliminary tests was Leiden, an algorithm proposed by [13] and introduced for single-cell analysis by [8].

We began by constructing a k-nearest neighbor (kNN) graph. This graph helps determine the connectivity between data points based on their nearest neighbours. By setting the number of nearest neighbours, we defined the local neighborhood size for each data point. This step was

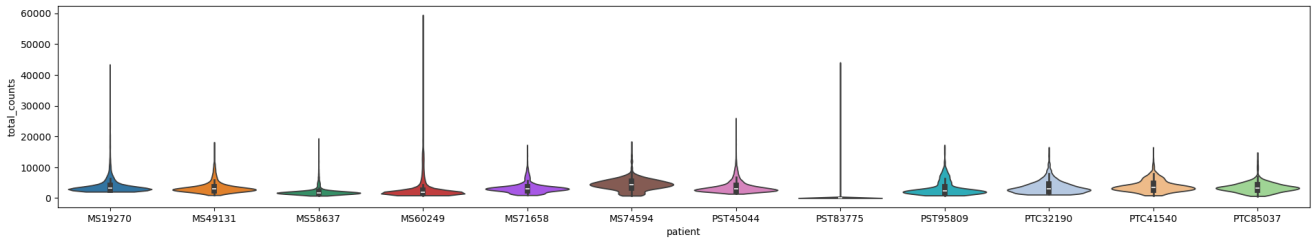


Figure 8: Violin plots illustrate the distribution of the total number of genes detected per CSF cell for each patient in the GSE138266 dataset. The dataset includes samples from 6 MS patients (identified by codes starting with MS) and 6 healthy controls (identified by codes starting with PST and PTC).

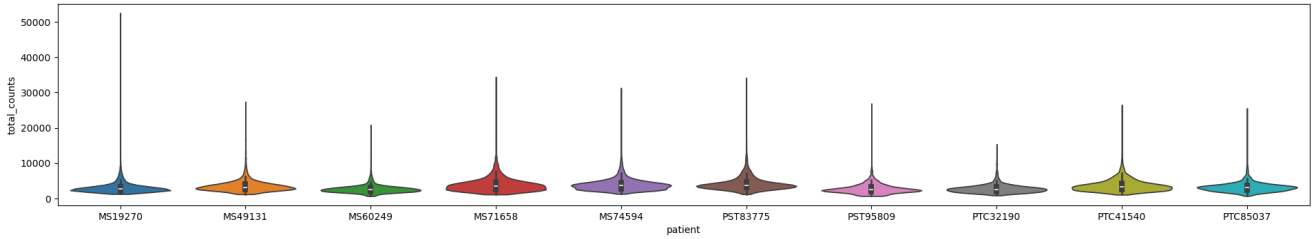


Figure 9: Violin plots illustrate the distribution of the total number of genes detected per CSF cell for each patient in the GSE138266 dataset. The dataset includes samples from 5 MS patients (identified by codes starting with MS) and 5 healthy controls (identified by codes starting with PST and PTC).

crucial for creating a network representation of the data, which formed the basis for subsequent clustering analysis.

Then, we applied the Uniform Manifold Approximation and Projection (UMAP) technique to visualize the high-dimensional data in a lower-dimensional space. UMAP is a manifold learning algorithm that preserves the global structure of the data while reducing its dimensionality. By embedding the data into a lower-dimensional space, UMAP facilitates the identification of clusters and patterns in the data. Our choice of this method is due to its speed and ability to preserve the underlying data topology. The implementation we used is presented in [10].

Finally, we employed the Leiden algorithm to partition the data into distinct clusters. The Leiden algorithm is a community detection method that identifies densely connected subgroups within a graph.

From the tests we conducted, we observed that by adjusting the resolution parameter, which controls the granularity of the clustering, higher values led to more fine-grained clusters. This step allowed us to identify and characterize cell subpopulations based on their transcriptional profiles, providing insights into the underlying biological processes. These observations aided us in selecting the parameters to use.

The results of the grid search are reported in the Table 3 below. The range of parameters was selected based on preliminary tests, and given the computational time required for each test (all run until convergence without early stopping) and the already satisfactory results, we decided to not explore further parameter configurations.

To assess the quality of clustering and identify the optimal parameter combination, we evaluated several metrics: the Silhouette score, to measure the cohesion and separation of clusters by assessing how similar each data point is to its own cluster compared to others, the Calinski-Harabasz score, to quantify the compactness and separation between clusters, and the Davies-

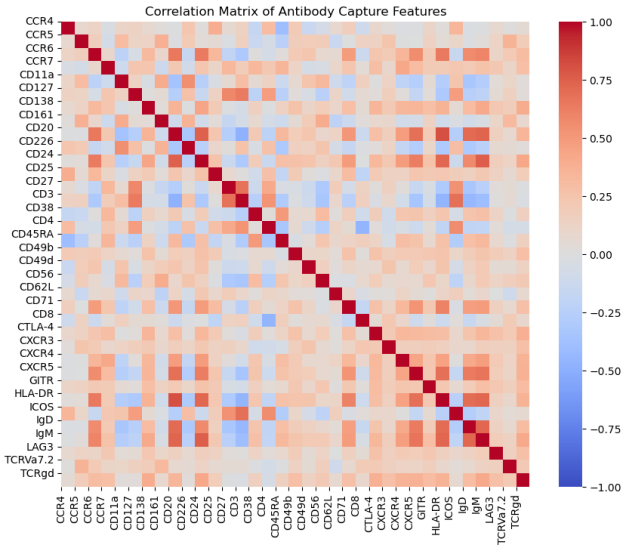


Figure 10: Correlation matrix of antibody features, indicating the degree of correlation among the 35 antibody features used in the study, providing a comprehensive view of the relationships between various antibody capture features. Correlation matrix showcases a wide range of correlation values, from strongly positive (red) to strongly negative (blue).

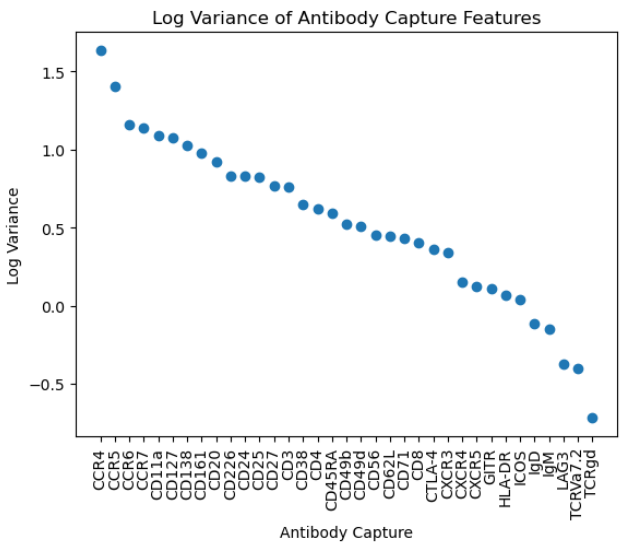


Figure 11: Logarithmic variance of the antibody capture features. The figure highlights the variance of each feature. The features CCR4, CCR5, and CCR6 show the highest logarithmic variances, indicating significant variability in their expression levels across samples. Conversely, features such as TCRV7.2, TCRgd, and LAG3 exhibit very low logarithmic variances, suggesting minimal variability in their expression levels.

Bouldin score, to evaluate the average similarity of each cluster with its most similar cluster. These metrics offer distinct insights into different aspects of clustering performance.

The evaluation scores for clustering are presented in Table 3, highlighting the best parameter combination, with respect to the Silhouette score.

neighbors	resolution	Silhouette	Calinski−Harabasz	Davies−Bouldin
5	0.5	0.0997	3937.4035	1.9732
5	0.8	0.0636	3114.0382	2.1261
5	1.0	0.0684	2943.2850	2.2308
10	**0.5**	**0.1255**	**4947.5156**	**2.0338**
10	0.8	0.1082	4214.3011	2.2261
10	1.0	0.0829	3850.6357	2.1951

Table 3: Performance evaluation of the Leiden clustering algorithm applied to the selected five principal components for antibody data analysis. The table displays clustering performance metrics for various parameter combinations, encompassing Silhouette score, Calinski−Harabasz score, and Davies−Bouldin score.

The plot of clusters in the two principal component UMAP space of the 16 clusters defined by the selected algorithm is depicted in Figure 12.

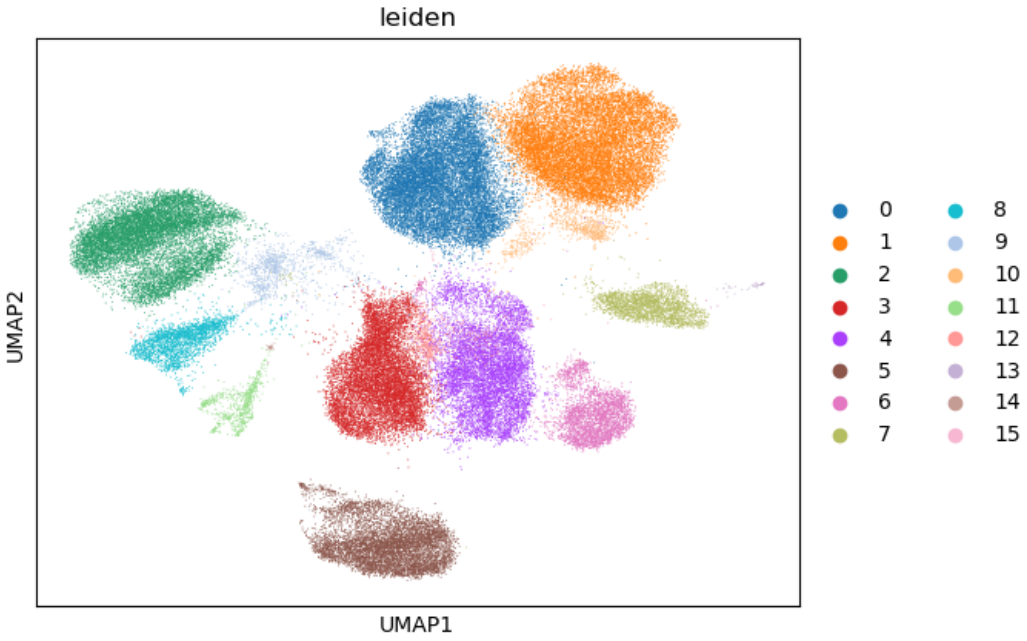


Figure 12: Figure showing the distribution of clusters in the two principal component UMAP space. The clusters are defined by the Leiden algorithm applied to antibody features.

3.2 Cell Type Labels

We inferred cell type labels using CellTypist[6], a library that allows for automatic cell annotation based on transcriptomic data.

We utilized the `Immune_All_High` pre-trained model from the library, which was trained using a collection of data from immune populations combined from 20 tissues of 18 studies [3].

The input to the model consisted of transcriptomic data containing gene counts for all cells in the GSE239626 dataset. Additionally, the clustering labels obtained with the Leiden algorithm, as explained in Section 3.1, were also fed into the model. This integration enhances annotation accuracy by incorporating information from the preliminary clustering obtained using antibody features.

Final annotations were determined through a majority voting mechanism, further enhancing the robustness of predictions.

Figure 13 shows the bar plot of label counts. We observed that the majority of cells were assigned the label T cells (70.9128% of the dataset), followed by B cells (17.9460%) and ILC (9.9783%). Other cell types had much lower proportions.

For each cell type label, we selected the top 5 markers (genes) based on their expression levels and visualized them in Figures 14 using dot plots. In these plots, each dot represents two values: the mean expression level within each category (indicated by color) and the fraction of cells expressing the gene in that category (indicated by the size of the dot). This visualization approach allows us to identify the most discriminative markers for each cell type as well as those shared by multiple cell lines.

From the dot plot, we observe that the markers most relevant for T cells are also broadly expressed by Cycling cells, with the exception of the genes IL7R and CD3G, which are less expressed in cycling cells. These genes are thus the discriminating markers between the two. In all other cell types, the markers characteristic of T cells are either not expressed or are expressed

[6] `https://www.celltypist.org/`

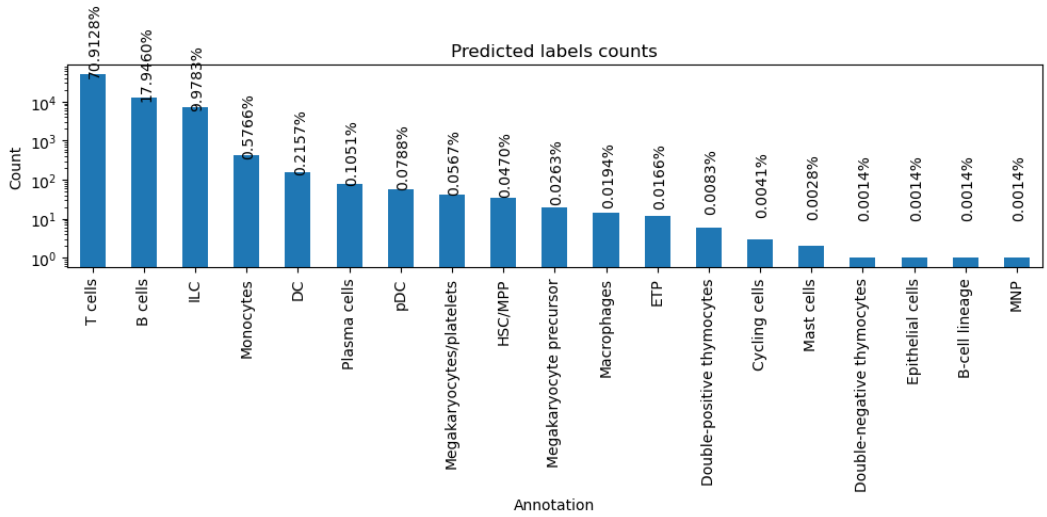


Figure 13: Barplot illustrating the distribution of predicted cell type labels across the GSE239626 dataset. The majority of cells were assigned the label T cells (70.91% of the dataset), followed by B cells (17.95%) and ILC (9.98%). Other cell types had much lower representation.

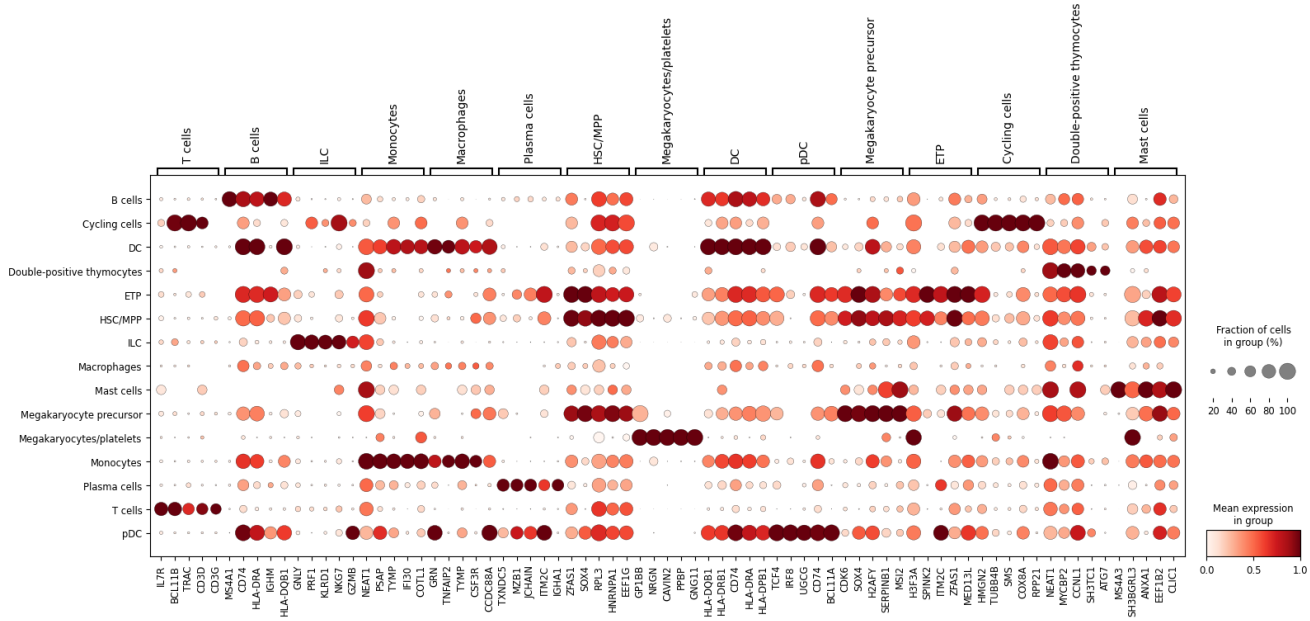


Figure 14: Dot plot illustrating the top 5 marker genes for each predicted cell type label. Each dot represents two values: mean expression within each category (visualized by color) and the fraction of cells expressing the gene in the category (visualized by the size of the dot).

at very low levels, indicating that this cell type is well-defined by its specific markers. For B cells, their characteristic genes are partially expressed in dendritic cells (DC) and plasmacytoid dendritic cells (pDC). Additionally, the gene CD74 is a characteristic marker not only for B cells and DCs but also for monocytes, while the gene HLA-DRA is highly expressed across several other cell types as well. In the case of monocytes, we notice that their characteristic genes are also expressed in CD cells. For plasma cells, their characteristic genes are also expressed in pDCs, except for the gene IGHA1, which is unique to plasma cells. Lastly, we observe an overlap of highly expressed genes between DCs and pDCs, indicating some commonality in their marker expression profiles.

3.3 Cell Type Labels Transfer

Before starting our differential expression analysis, we undertook the task of transferring cell type labels extracted from the GSE239626 dataset to the GSE194078 and GSE138266 datasets. This process ensured label consistency across all datasets, facilitating the creation of a unified dataset for further analysis.

To do this, we integrate the datasets by first log-transforming all the original data in order to ensure comparability. Then, we perform dimensionality reduction and neighbourhood graph construction on the reference dataset GSE239626, using as parameters the ones selected for cell type extraction, see Table 3. The new datasets are integrated with the reference dataset using ScanPy built-in function `Ingest`, that transfer cell type labels from the reference to the new datasets. This step ensures that the new datasets are annotated in a manner consistent with the reference dataset. The two datasets, GSE194078 and GSE138266 respectively, are then concatenated with the reference dataset, retaining the batch information to distinguish between the reference and new datasets.

In order to reduce batch effect in our analysis, we use BBKNN (Batch Balanced K-Nearest Neighbours), proposed by [11]. The BBKNN algorithm corrects batch effects in single-cell RNA-Seq data by modifying the construction of the k-nearest neighbours (kNN) graph. It first identifies k-nearest neighbours for each cell within its batch using an approximate nearest neighbor search. Then, it balances these neighbours by incorporating a fixed number of neighbours from each other batch.

Figures 15 and 16 illustrate the representations of the merged dataset comprising the reference dataset GSE239626 and the two additional datasets, GSE138266 and GSE194078, respectively, in the two principal component UMAP space. In both figures, distinct cell types are highlighted with different colors.

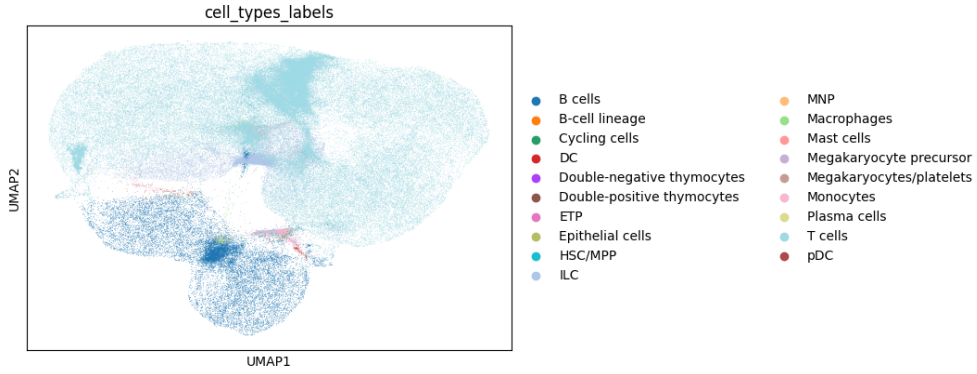


Figure 15: Representations of the merged dataset comprising the reference dataset GSE239626 with the additional datasets GSE138266 in the two principal component UMAP space. In the figures, distinct cell types are highlighted with different colors.

The distribution of predicted labels in the GSE194078 and GSE138266 datasets, relative to the total, is represented in the barplots in Figures 21 and 22 in Appendix, respectively.

The barplots indicate that the most common cell type in both datasets is T cells (75.6592% in GSE138266 and 82.2855% in GSE194078), followed by B cells (10.2634% and 11.6746%, respectively). These results align with the labels inferred from the GSE239626 dataset used as a reference.

ILC cells, the third most prevalent cell type in the GSE239626 dataset, rank third in GSE138266 (7.4664%) and fourth in GSE194078 (2.5379%, after monocytes with 2.6392%).

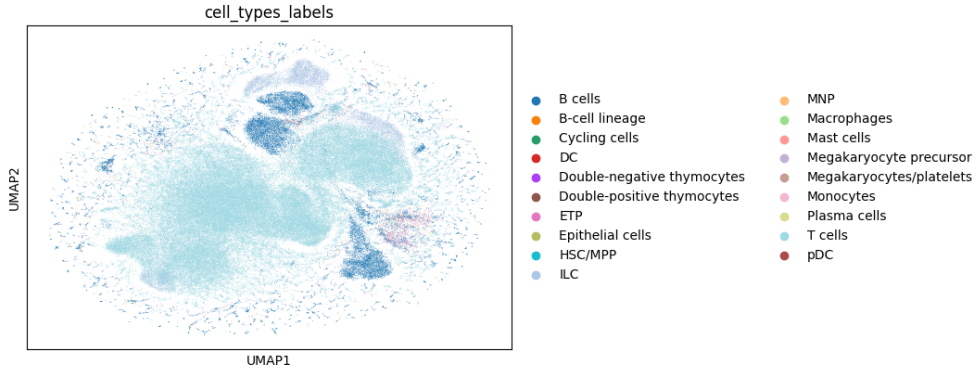


Figure 16: Representations of the merged dataset comprising the reference dataset GSE239626 with the additional datasets GSE194078 in the two principal component UMAP space. In the figures, distinct cell types are highlighted with different colors.

In Figure 17 we visualized the top 20 highest expressed genes in the dataset. As shown, the MALT1 gene has the highest expression, with a significant percentage of total counts. This result was expected since MALT1 is the most expressed gene in the three source datasets used to generate ours. The large box and numerous outliers indicate variability in its expression levels across different cells, a pattern observed in the boxplots for all top 10 expressed genes. Several mitochondrial genes, such as MT-CO3 and MT-CO1, also appear among the top expressed, indicating substantial mitochondrial activity. The presence of these genes underscores the importance of mitochondrial functions in the cellular processes represented in our dataset. Some ribosomal protein genes, such as RPS27 and HSPA5, are also highly expressed. Notably, the distribution of HSPA5 features very long whiskers, indicating significant variability in expression levels across different cells. This suggests that HSPA5 expression is highly heterogeneous within the dataset, reflecting diverse cellular conditions and states.

4 Differential Expression Analysis

In this section, we explain the methodology we followed to discover quantitative changes in expression levels between experimental groups. We treat cerebrospinal fluid (CSF) cells and peripheral blood mononuclear cells (PBMC) separately to ensure accurate analysis and comparison.

The cells used for this analysis include T cells, B cells, ILC, Monocytes, Plasma cells, HSC/MPP, and DC, as listed in the table. These cells are available in varying numbers in both PBMC and CSF samples and also show distinct numerical differences between healthy and diseased patient samples.

The cardinality of each cell type with respect to PBMC or CSF and HC or MS is reported in Table 4 below and visualized in Figures 20 in Appendix.

We note that the number of cells, for each type, reported in MS samples is generally higher than in HC samples. This is particularly evident for T cells in the blood, where the number of cells in the dataset for MS patients is almost double that of HC . This discrepancy is also due to the imbalance in the number of samples between the MS and HC classes in our dataset.

4.1 Most Expressed Genes

To compute and rank differentially expressed genes between groups, specifically comparing MS (Multiple Sclerosis) versus HC (Healthy Control), we employed the `rank_genes_groups` method

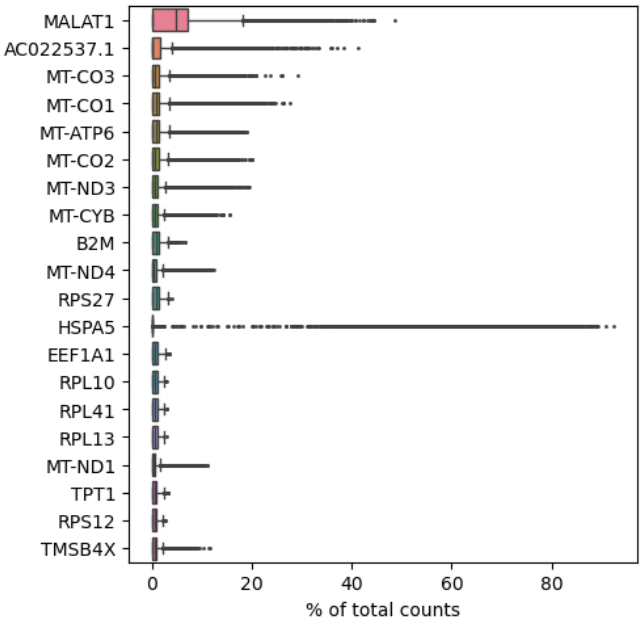


Figure 17: Box plot illustrating the distribution of gene expression levels for the top 20 genes in our transcriptomic dataset. The MALT1 gene shows the highest expression with significant variability. The presence of several mitochondrial genes, such as MT-CO3 and MT-CO1, highlights substantial mitochondrial activity. Additionally, some ribosomal protein genes, like RPS27 and HSPA5, are highly expressed. Notably, the distribution of HSPA5 features long whiskers, indicating significant variability in its expression levels across different cells.

Cell Type	**PBMC MS**	**PBMC HC**	**CSF MS**	**CSF HC**
T cells	84479	48474	31911	28825
B cells	16866	7020	2905	6088
ILC	10672	3104	889	471
Monocytes	2003	2710	579	1357
Plasma cells	809	570	172	461
HSC/MPP	81	14	2	24
DC	187	38	15	29

Table 4: Cell counts across different conditions and cell types. PBMC MS and PBMC HC represent peripheral blood mononuclear cells in multiple sclerosis patients and healthy controls, respectively. CSF MS and CSF HC represent cerebrospinal fluid cells in multiple sclerosis patients and healthy controls, respectively. The most frequent cells are T cells, followed by B cells. The number of cells varies between different conditions, but the proportions of increase or decrease remain consistent across conditions.

from the ScanPy library.

Initially, we explored the `t-test` method suggested by the library to see if we could achieve satisfactory results while taking advantage of its computational speed. However, `t-test` requires normally distributed data.

To overcome this limitation, we explored another method: the `wilcoxon` method, which uses the Wilcoxon rank-sum test. This method better matches our data assumptions, as it does not require the data to follow a specific distribution.

Despite the computational speed of the t-test, we decided to use the Wilcoxon rank-sum test

based on the results obtained.

To visualize the results, we initially used a p-value threshold of 0.01. From this preliminary exploratory analysis, we decided to notinclude HSC/MPP cells in our study, as we were unable to identify differentially expressed genes between the two groups with a low p-value. Additionally, the number of samples for these cells was low. A similar decision was made for DC cells in the CSF.

The results of the most highly expressed genes identified for each cell type (with a p-value $<$ 0.01) in both PBMC and CSF cells are visualized using Venn diagrams in Figure 24 in Appendix.

The heatmaps in Figure 18 illustrate the expression levels of the 20 most significant genes across various cell types in MS (Multiple Sclerosis) versus HC (Healthy Control) samples, both in PBMC and CSF.

The heatmaps reveal that T cells, B cells, monocytes, and plasma cells, particularly in PBMC samples, show the most significant differences in gene expression between MS and HC. These cell types may play a crucial role in the pathogenesis of MS. ILCs and DCs also show differential expression, but the differences are generally less pronounced compared to other cell types. The PBMC compartment appears to have more marked differences overall, suggesting that peripheral blood may be a more informative source for identifying gene expression changes associated with MS.

To determine which genes were deemed significant and thus relevant for our analyses, we utilized a threshold value for the adjusted p-value of 10^{-30} and a log fold change of 0.5.

The volcano plots reported in Figure 26 in Appendix visually illustrate genes that surpass these significance criteria, facilitating the identification of those showing the most pronounced differences in expression between the groups.

5 Differential Expression Results

In Figures 27 and 28 in Appendix, barplots are presented to visualize the number of upregulated and downregulated genes for each cell type in the PBMC and CSF cell populations, respectively.

The number of selected upregulated and downregulated genes for each cell type is reported in Table 5 for PBMC and CSF cells.

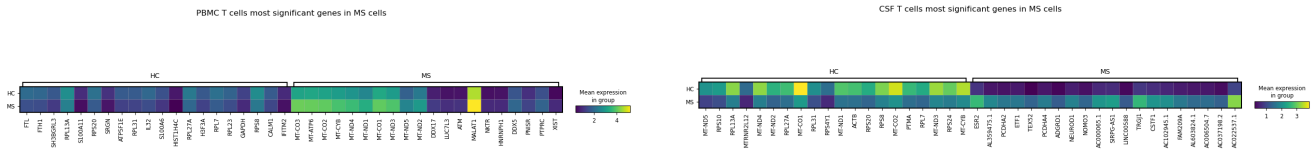
	PBMC		**CSF**	
Cell Type	**Downregulated**	**Upregulated**	**Downregulated**	**Upregulated**
B cells	3309	222	2907	1008
ILC	117	99	287	165
Monocytes	891	174	1610	853
Plasma cells	1513	205	108	19
T cells	543	271	3436	1395

Table 5: Number of upregulated and downregulated genes for each cell type in PBMC and CSF samples.

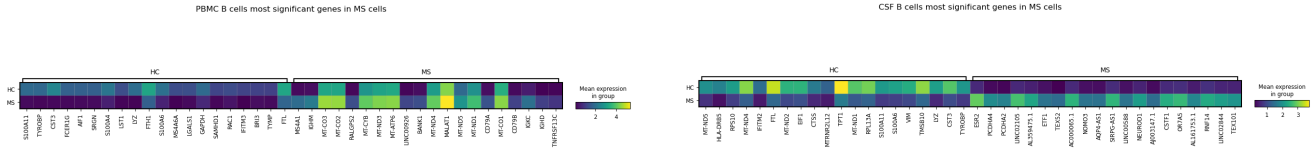
It can be observed that the number of downregulated genes is much higher than that of upregulated genes across all cell types.

Additionally, by using the same criteria for selecting the most relevant genes for each cell type, as outlined in Section 4.1, we noted that the number of relevant genes in T cells is much higher than in other cell types, with the exception of downregulated genes in B cells in PBMC.

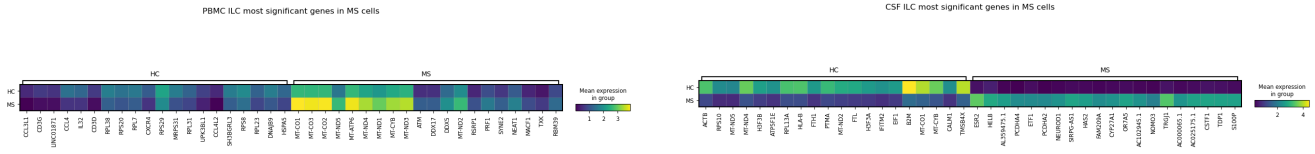
To conduct a deeper analysis of significant genes, we examined the upregulated and downregulated gene expressions in cells from multiple sclerosis (MS) patients compared to healthy controls



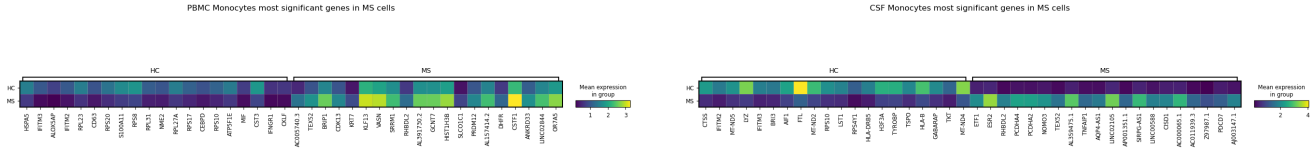
(a) T cells



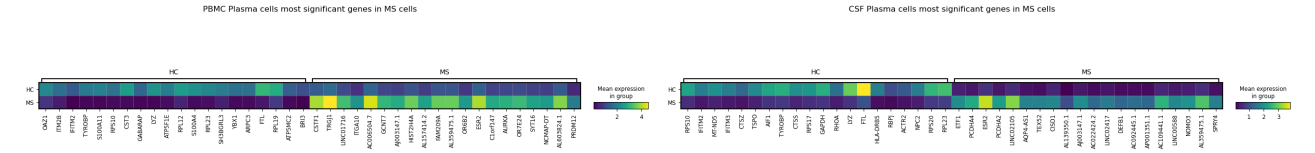
(b) B cells



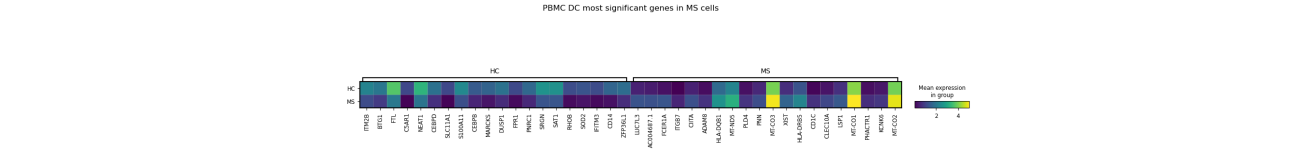
(c) Innate Lymphoid Cells (ILCs)



(d) Monocytes



(e) Plasma cells



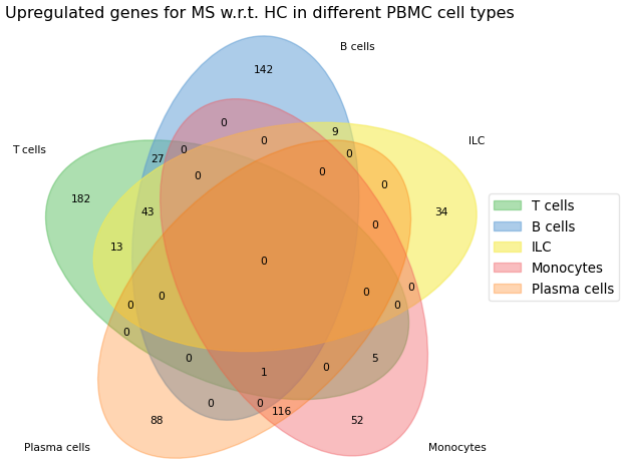
(f) Dendritic Cells (DCs)

Figure 18: Heatmaps depicting the 20 most significant genes in various cell types of MS (Multiple Sclerosis) versus HC (Healthy Control) samples in both PBMC and CSF. Each heatmap shows the mean expression levels of the top differentially expressed genes in T cells, B cells, ILCs, monocytes, plasma cells, and DCs, categorized into MS and HC groups. The color gradient represents the expression levels, with darker colors indicating lower expression and lighter colors indicating higher expression.

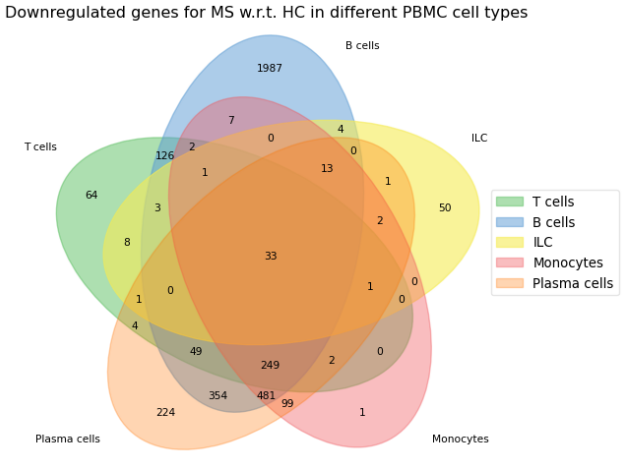
(HC) in both Peripheral Blood Mononuclear Cells (PBMC) and Cerebrospinal Fluid Cells (CSFs) populations.

Venn diagrams illustrating these comparisons are presented in Figures 19a, 19b, 19c, and 19d below.

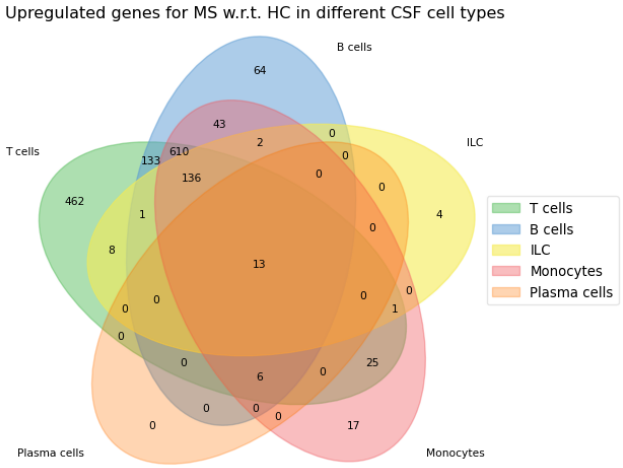
The first Venn diagram in Figures 19a, illustrates the upregulated genes observed in the PBMCs of MS patients compared to HCs. Among the various cell types, T cells exhibit the highest number of unique upregulated genes, totaling 182, while B cells show a significant count of 142 unique upregulated genes. Interestingly, 71 upregulated genes are commonly expressed by both



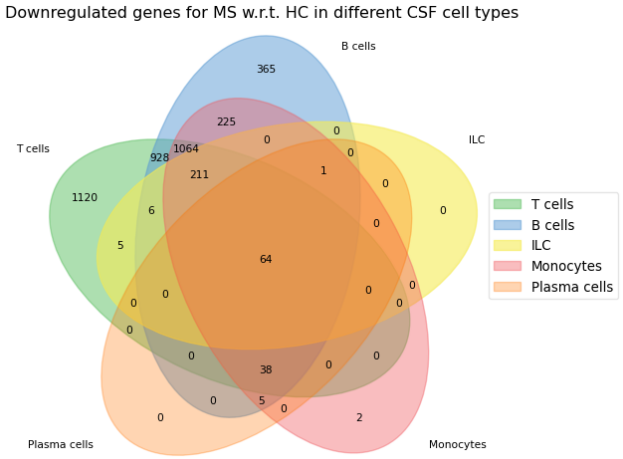
(a) Venn diagram illustrates the upregulated genes observed in the PBMCs of MS patients compared to HCs.



(b) Venn diagram illustrates the downregulated genes observed in the PBMCs of MS patients compared to HCs.



(c) Venn diagram illustrates the upregulated genes observed in the CSFs of MS patients compared to HCs.



(d) Venn diagram illustrates the downregulated genes observed in the CSFs of MS patients compared to HCs.

Figure 19: Venn diagrams illustrates the upregulated and downregulated genes observed in MS patients compared to HCs.

T cells and B cells, with 43 of these also occurring in ILC. Additionally, a notable overlap of 116 genes is observed between Plasma cells and Monocytes, suggesting shared upregulated pathways between these cell types. These overlapping patterns underscore similarities in their response to MS. Notably, no gene is found to be common across all five PBMC cell types, indicating highly specific upregulation profiles depending on the cell type.

In Figures 19b, the downregulated genes in PBMCs are represented. B cells exhibit the highest number of unique downregulated genes, with 1987 genes affected, contributing to the highest overall count of downregulated genes, totaling 3309 across all cell types. Among these, a notable overlap of 126 genes is observed between only T cells and B cells, indicating shared suppression mechanisms. Additionally, 354 downregulated genes are shared only between plasma cells and B cells, suggesting some commonality in their response to MS. Moreover, an overlap of 481 genes is evident among B cells, Monocytes, and Plasma cells, highlighting potential interconnected pathways affected by downregulation.

We then directed our analysis towards Cerebrospinal Fluid (CSF) cell populations, with a focus

on upregulated genes. As depicted in Figures 19c, the analysis highlights the upregulated genes across five distinct CSF cell types: T cells, B cells, Innate Lymphoid Cells (ILC), Monocytes, and Plasma cells. Particularly noteworthy is the observation that T cells display the highest number of unique upregulated genes, totaling 462. This suggests a substantial alteration in gene expression specific to T cells within the CSF of MS patients. Furthermore, a significant overlap of 610 genes shared between T cells, B cells, and Monocytes indicates potential common pathways or functions being upregulated across these cell types. Additionally, smaller overlaps exist among other cell types, with 13 genes being common to all five cell types, highlighting potential universal mechanisms involved in MS pathology.

Moving on to the downregulated genes in CSF cell types, the final Venn diagram in Figures 19d illustrates that T cells exhibit the highest number of unique downregulated genes, totaling 1120. This extensive downregulation points to significant suppression of gene activity in T cells within the CSF of MS patients. The substantial overlap of 1064 genes shared between T cells, B cells, and Monocytes again suggests shared pathways or cellular responses being suppressed. Notably, T cells and B cells emerge as the cell types with the highest number of shared genes detected, totaling 2311, of which 928 are uniquely shared between these two cell types. Furthermore, 64 genes are downregulated across all five cell types, indicating a broad suppression of certain functions or pathways in MS.

In summary, the data indicate that T cells and B cells, especially in the CSF, exhibit extensive differences in gene expression, both in terms of upregulation and downregulation.

The genes that are upregulated and downregulated in common across all analyzed cell types (T cells, B cells, Monocytes, Plasma cells, and ILC) are reported in Table 11 in Appendix.

6 Comparative Analysis in B Cells

We compared the up and downregulated genes in B cells, which are differentially expressed between Multiple Sclerosis (MS) patients and healthy controls (HC), across the GSE173787 dataset and our transcriptomic dataset for PBMC cells.

Using the same threshold for log fold change as applied in the differential expression analysis of our transcriptomic dataset (logFC $<$ -0.5 for determining downregulated genes and logFC $>$ 0.5 for upregulated genes), we selected differentially expressed genes in MS and HC for the GSE173787 dataset.

It is important to note that we used a p-value threshold set to 10^{-30} to determine the significance of genes within our transcriptomic dataset. However, for this dataset, we could not employ such a stringent threshold because the reported p-value values for the genes in this dataset had a lower bound of 10^{-13}, as reported in table 2 in Section 2.4. Therefore, we considered all genes that met the criteria based on logFC, regardless of the reported p-value value.

In the GSE173787 dataset, 417 upregulated genes and 130 downregulated genes was identified for B cells. In our dataset, we found 222 upregulated genes and 3309 downregulated genes in the same cell type.

Between the two datasets, only 3 upregulated genes and 14 downregulated genes were shared. The shared upregulated genes are IGLC3, IGLC2, and IGKC, while the shared downregulated genes include CCDC28A, OGFRL1, BRI3, HSPB1, IFNGR1, MID1IP1, CSNK1E, EMC6, DUSP6, NDUFV2, DHRS3, CLEC2B, CMTM3, and BEX4.

Venn diagram in Figure 29 in Appendix illustrating the comparison of up and downregulated genes between the GSE173787 dataset and our transcriptomic dataset.

7 Limitations and Future works

The scarcity of comprehensive data at our disposal and the absence of a representative statistical sample can undermine the validity and generalizability of our findings. Limited data availability may impede the comprehensiveness and representativeness of the information utilized, potentially introducing bias to our results. These limitations underscore the importance of cautious interpretation and highlight the need for further research to address data gaps and enhance the robustness of future analyses.

8 Conclusions

In conclusion, our analysis has identified genes that are differentially expressed between MS patients and healthy controls, with a notable prevalence of downregulated genes in MS patients compared to HC. These differences are particularly pronounced in the CSF of MS patients, underscoring the importance of further investigation in this direction. Moreover, our findings reveal significant disparities in gene expression profiles between T cells and B cells in the Cerebrospinal Fluid Cells of MS patients, implicating intricate alterations in cellular functions and signaling pathways associated with Multiple Sclerosis pathology.

Additionally, the comparative analysis of differential gene expression in B cells between the markers found in the GSE173787 dataset and our transcriptomic dataset for PBMC cells underscores the variability in gene expression patterns across different studies. While some genes exhibit consistency in their dysregulation between MS patients and healthy controls, there are notable discrepancies, highlighting the influence of various factors such as sample heterogeneity, experimental techniques, and data processing methods.

Code availability

The exploratory data analysis performed on the datasets and the preprocessing steps necessary to make the data usable are available on GitHub at `https://github.com/multi-omics-integration-for-MS/MS-GEO-Datasets`. The scripts used to perform data analysis and obtain results are available on GitHub at `https://github.com/multi-omics-integration-for-MS/multi-omics-integration`.

References

[1] Sergio E Baranzini and Jorge R Oksenberg. The genetics of multiple sclerosis: from 0 to 200 in 50 years. *Trends in genetics*, 33(12):960–970, 2017.

[2] Ruth Dobson and Gavin Giovannoni. Multiple sclerosis—a review. *European Journal of Neurology*, 26(1):27–40, 2019.

[3] C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

[4] Manon Galoppin, Manon Rival, Anaïs Louis, Saniya Kari, Sasha Soldati, Britta Engelhardt, Anne Astier, Philippe Marin, and Eric Thouvenot. Cite-seq reveals inhibition of nf-kb path-

way in b cells from vitamin d-treated multiple sclerosis patients. *bioRxiv*, pages 2023–09, 2023.

[5] Carol Chase Huizar, Itay Raphael, and Thomas G Forsthuber. Genomic, proteomic, and systems biology approaches in biomarker discovery for multiple sclerosis. *Cellular immunology*, 358:104219, 2020.

[6] Vanitha A Jagannath, Graziella Filippini, Israel Junior Borges do Nascimento, Carlo Di Pietrantonj, Edward W Robak, and Liz Whamond. Vitamin d for the management of multiple sclerosis. *Cochrane database of systematic reviews*, (9), 2018.

[7] Junho Kang, Moonhang Kim, Da-Young Yoon, Woo-Seok Kim, Seok-Jin Choi, Young-Nam Kwon, Won-Seok Kim, Sung-Hye Park, Jung-Joon Sung, Myungsun Park, et al. Axl+ siglec6+ dendritic cells in cerebrospinal fluid and brain tissues of patients with autoimmune inflammatory demyelinating disease of cns. *Clinical Immunology*, 253:109686, 2023.

[8] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[9] Qin Ma, Stacy J Caillier, Shaun Muzic, University of California San Francisco MS-EPIC Team, Michael R Wilson, Roland G Henry, Bruce AC Cree, Stephen L Hauser, Alessandro Didonna, and Jorge R Oksenberg. Specific hypomethylation programs underpin b cell activation in early multiple sclerosis. *Proceedings of the National Academy of Sciences*, 118(51):e2111920118, 2021.

[10] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[11] Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, 08 2019.

[12] David Schafflick, Chenling A Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G Meuth, Tanja Kuhlmann, et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature communications*, 11(1):247, 2020.

[13] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019.

Appendix

GEO accession	GSE239626
Title	CITE-seq reveals inhibition of NF-kB pathway in B cells from vitamin D-treated multiple sclerosis patients
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Vitamin D deficiency is a risk factor for multiple sclerosis (MS) and is correlated with disease activity and progression. Vitamin D treatment has emerged as potentially protective, even though results from randomized controlled trials are conflicting. Here, we used single-cell RNA-seq (scRNA-seq) coupled with barcoded antibodies targeting surface markers (CITE-seq) to discover candidate genes and pathways regulated in PBMC subpopulations from MS patients treated with high-dose vitamin D (n=5) or placebo (n=5). Best candidates were combined with genes involved in immune function and vitamin D metabolism for validation in a new cohort (n=8 in each group) by high-throughput qPCR (HT-qPCR) in naive CD4, Th1, Th17, Treg, naive CD8, memory and naive B cells, and MAIT cells, sorted by FACS. CITE-seq showed no significant changes in the proportions of these subpopulations in vitamin D-treated patients. Of the 92 candidate genes revealed by CITE-seq, differential expression of five genes (UXT, SNRPN, SUB1, GNLY and KLF6) was validated by HT-qPCR. CITE-seq also revealed the regulation of several pathways by vitamin D in naive and memory B cells, including MAPK, TLR and interleukin pathways, that contribute to counteract EBV-induced resistance to apoptosis through inhibition of the NF-kB pathway.
Overall design	Five MS patients treated with high dose vitamin D (100,000 IU every 2 weeks for 3 months) and 5-matched MS patients treated with placebo were selected from the D-lay MS cohort (NCT01817166). Blood samples of study participants of the D-lay MS study had been previously collected in EDTA-containing tubes at D0 and M3 of treatment (placebo or vitamin D). PBMCs had been isolated by Ficoll® gradient and cryopreserved in liquid nitrogen using 4% human albumin solution + 10% DMSO medium in concentration of 5x106 cells per ml. Cells were labeled with CITE-seq antibodies and live cells were isolated by fluorescence-activated cell sorting (FACS).
Status	Public on October 5, 2023

Table 6: Description of Dataset GSE239626

GEO accession	GSE194078
Title	AXL+SIGLEC6+ dendritic cells in cerebrospinal fluid and brain tissues of patients with autoimmune inflammatory demyelinating disease of CNS
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Inflammatory demyelinating disease of the CNS (IDD) is a heterogeneous group of autoimmune diseases, and multiple sclerosis is the most common type. Dendritic cells (DCs), major antigen-presenting cells, have been proposed to play a central role in the pathogenesis of IDD. The AXL+SIGLEC6+ DC (ASDC) has been only recently identified in humans and has a high capability of T cell activation. Nevertheless, its contribution to CNS autoimmunity remains still obscure. Here, we aimed to identify the ASDC in diverse sample types from IDD patients and experimental autoimmune encephalomyelitis (EAE). A detailed analysis of DC subpopulations using single-cell transcriptomics for the paired cerebrospinal fluid (CSF) and blood samples of IDD patients (total n = 9) revealed that three subtypes of DCs (ASDCs, ACY3+ DCs, and LAMP3+ DCs) were overrepresented in CSF compared with their paired blood. Among these DCs, ASDCs were also more abundant in CSF of IDD patients than in controls, manifesting poly-adhesional and stimulatory characteristics. In the brain biopsied tissues of IDD patients, obtained at the acute attack of disease, ASDC were also frequently found in close contact with T cells. Lastly, the frequency of ASDC was found to be temporally more abundant in acute attack of disease both in CSF samples of IDD patients and in tissues of EAE, an animal model for CNS autoimmunity. Our analysis suggests that the ASDC might be involved in the pathogenesis of CNS autoimmunity.
Overall design	We have generated a single-cell transcriptome dataset for paired cerebrospinal fluid and PBMC samples from 9 Inflammatory demyelinating disease patients and 2 healthy controls without inflammatory disease of the central nervous system.
Status	Public on Jul 10, 2023

Table 7: Description of Dataset GSE194078

GEO accession	GSE138266
Title	Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Cerebrospinal fluid (CSF) protects the central nervous system (CNS) and analyzing CSF aids the diagnosis of CNS diseases, but our understanding of CSF leukocytes remains superficial. Here, using single cell transcriptomics, we identified a specific border-associated composition and transcriptome of CSF leukocytes. Multiple sclerosis (MS) – an autoimmune disease of the CNS – increased transcriptional diversity in blood, but increased cell type diversity in CSF including a higher abundance of cytotoxic phenotype T helper cells. A new analytical approach, named cell set enrichment analysis (CSEA) identified a cluster-independent increase of follicular T helper (TFH) cells potentially driving the known expansion of B lineage cells in the CSF in MS. In mice, TFH cells accordingly promoted B cell infiltration into the CNS and the severity of MS animal models. Immune mechanisms in MS are thus highly compartmentalized and indicate ongoing local T/B cell interaction.
Overall design	5 vs. 5 Case-Control design for the single cell RNAseq experiment
Samples	Each donor provided 2 samples from the cerebrospinal fluid and PBMC
Status	Public on Dec 10, 2019

Table 8: Description of Dataset GSE138266

GEO accession	GSE173787
Title	DNA methylation profiles of four immune cell types from MS patients and healthy controls
Organism	Homo sapiens
Experiment type	Methylation profiling by high throughput sequencing
Summary	How epigenetic changes contribute to MS pathogenesis is still poorly understood. Therefore, we conducted a comprehensive analysis of genome-wide DNA methylation patterns in four immune cell populations isolated from MS patients at clinical disease onset. We also performed parallel transcriptome analysis in B cells to better understand the functional consequences of the DNA methylation changes in MS.
Overall design	Four immune cell populations, namely CD4+ and CD8+T cells, CD14+ monocytes and CD19+ B cells, were FACS sorted from PBMC samples from a cohort of MS and HC subjects. Cell were then used for DNA extraction and DNA methylation profiling by bisulfite sequencing (BS-seq). The total RNA from B cell fractions were used for mRNA-seq.
Status	Public on Jan 18, 2022

Table 9: Description of Dataset GSE173787

Patients	**Count**	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**
GSM7669046	5491	6232.38	2989.81	311	4661.00	6413.00	7822.00	40142
GSM7669048	6018	5944.41	3116.15	373	3503.00	6161.50	7715.00	38287
GSM7669050	2583	3568.23	2742.93	328	1726.00	2401.00	5019.00	21706
GSM7669052	3143	4027.76	3075.89	355	1661.00	3541.00	5691.50	38932
GSM7669054	3579	3847.99	2852.84	503	1791.00	2657.00	5630.50	27790
GSM7669056	3884	3832.48	2804.96	317	1665.00	2561.00	5653.00	29256
GSM7669058	3809	3519.33	2597.80	474	1705.00	2580.00	4975.00	33825
GSM7669060	2822	3763.47	2779.34	446	1666.50	2715.00	5504.25	32205
GSM7669062	5006	4497.24	2829.46	500	2112.00	4227.50	6184.75	31364
GSM7669064	3160	3406.68	2590.65	432	1594.00	2350.50	4857.25	21941
GSM7669066	2945	4659.40	2449.54	351	2225.00	4911.00	6333.00	19381
GSM7669068	2709	4099.13	2743.76	452	1755.00	3462.00	6124.00	29950
GSM7669070	2956	2952.26	2614.63	449	1376.00	2005.50	3781.00	22534
GSM7669072	3715	4473.82	2579.04	335	1970.00	4637.00	6308.50	25461
GSM7669074	2543	4776.75	2584.43	449	2426.50	4989.00	6421.50	26231
GSM7669076	3946	2900.15	2420.58	408	1475.00	1938.50	3755.50	26415
GSM7669078	2779	3742.95	2326.52	408	1655.50	3580.00	5427.50	16089
GSM7669080	3964	3321.85	2401.90	329	1814.75	2390.00	4357.25	32057
GSM7669082	3423	3716.74	2342.02	318	1554.00	3574.00	5527.00	18089
GSM7669084	3842	3485.76	2343.37	400	1764.00	2441.00	5006.00	22022

Table 10: Summary statistics of total gene counts per cell for each of the 10 patients in the dataset GSE239626. The statistics include the number of observations (*Count*), average gene count (*Mean*), standard deviation (*Std*), minimum gene count (*Min*), first quartile (*25%*), median (*50%*), third quartile (*75%*), and maximum gene count (*Max*). Each patient has two unique IDs, shown in the *Patients* column, corresponding to samples collected at two different time points (initial and three months later). Each row represents a specific sample for the respective patient. Specifically, GSM7669046 and GSM7669048 represent the samples collected from the first patient at the initial time point and after three months, respectively; GSM7669050 and GSM7669052 from the second patient; Specifically, GSM7669046 and GSM7669048 represent the samples collected from the first patient at the initial time point and after three months, respectively; GSM7669050 and GSM7669052 from the second patient; GSM7669054 and GSM7669056 from the third patient; GSM7669058 and GSM7669060 from the fourth patient; GSM7669062 and GSM7669064 from the fifth patient; GSM7669066 and GSM7669068 from the sixth patient; GSM7669070 and GSM7669072 from the seventh patient; GSM7669074 and GSM7669076 from the eighth patient; GSM7669078 and GSM7669080 from the ninth patient; GSM7669082 and GSM7669084 from the tenth patient.

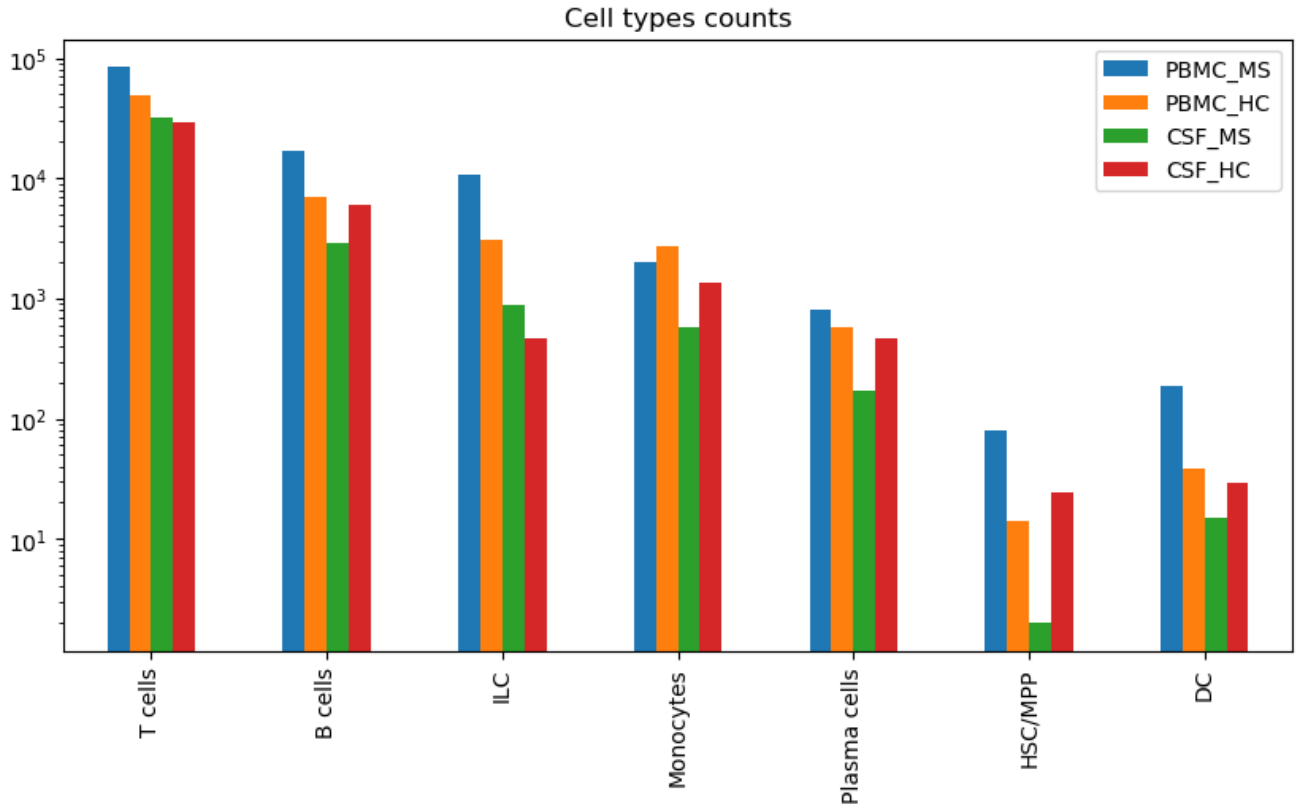


Figure 20: Bar plot showing the cardinality of each cell type with respect to PBMC or CSF and HC or MS samples. The cell types include T cells, B cells, ILC, Monocytes, Plasma cells, HSC/MPP, and DC. The y-axis is presented on a logarithmic scale to better visualize the differences in cell counts across the different conditions. Notably, the number of cells reported in MS samples is generally higher than in HC samples, with a particularly pronounced difference in T cells in the blood. This discrepancy reflects the imbalance in the number of samples between the MS and HC classes in our dataset.

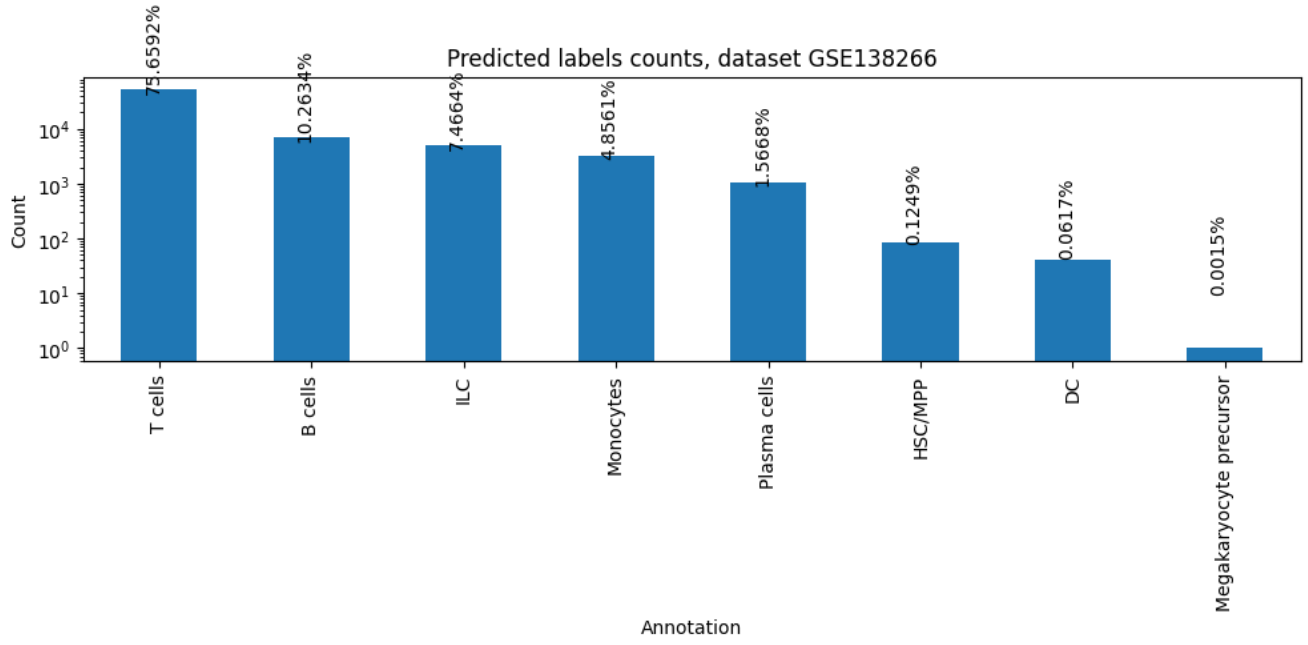


Figure 21: Bar chart plot showing the distribution of predicted cell types in the dataset GSE138266. T cells are the most abundant at 75.6592%, followed by B cells at 10.2634%. Other cell types are present in lower proportions, with Megakaryocyte precursors being the least frequent at 0.0015%

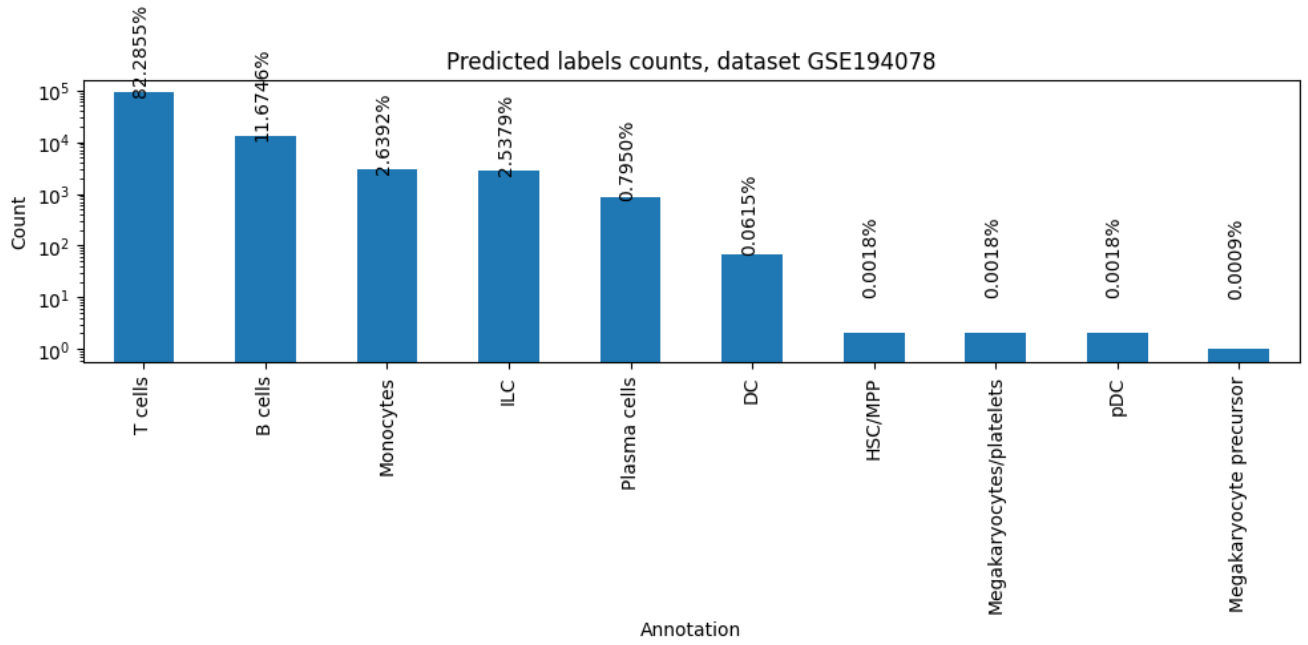
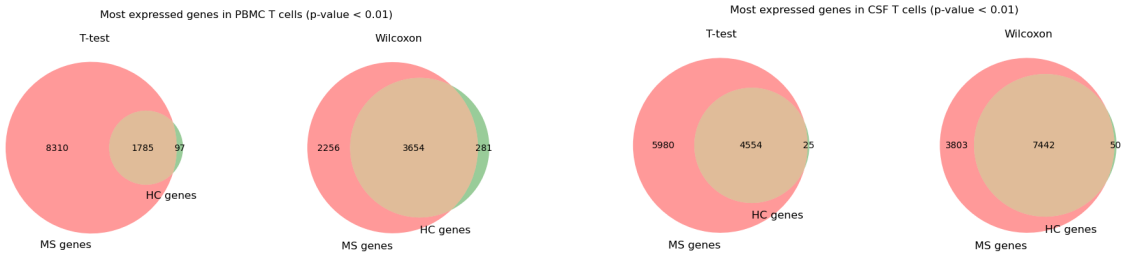
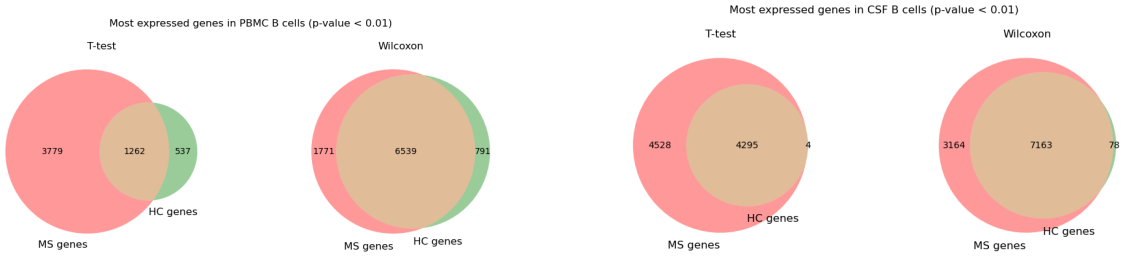


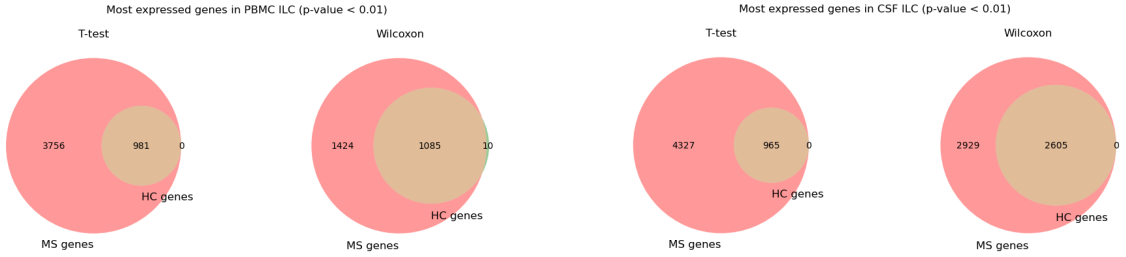
Figure 22: Bar chart plot showing the distribution of predicted cell types in the dataset GSE194078. T cells are the most abundant at 82.2855%, followed by B cells at 11.6746%. Other cell types are present in lower proportions, with Megakaryocyte precursors being the least frequent at 2.6392%



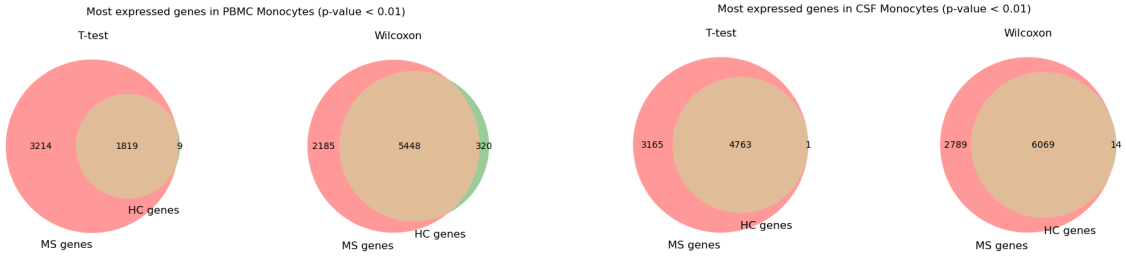
(a) T cells



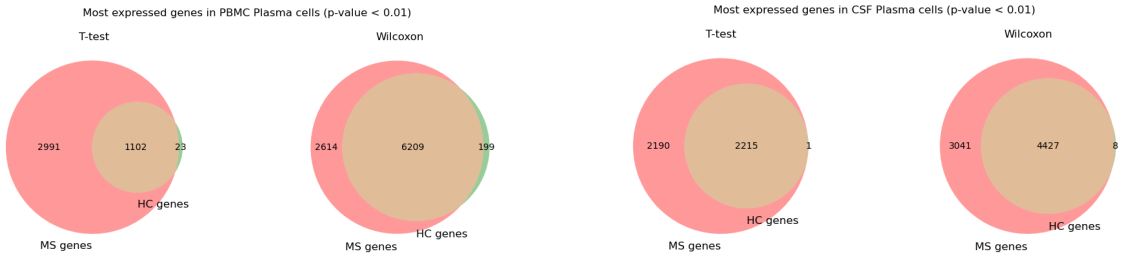
(b) B cells



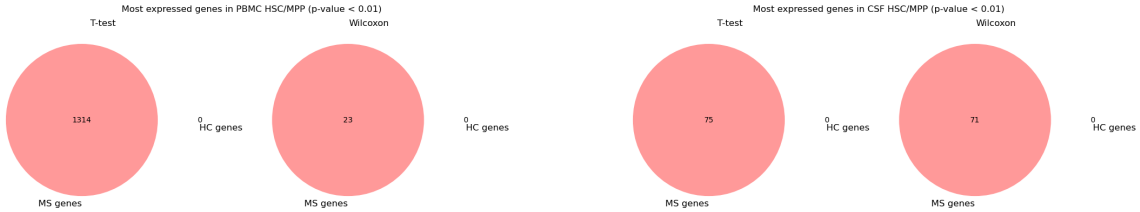
(c) Innate Lymphoid Cells (ILCs)



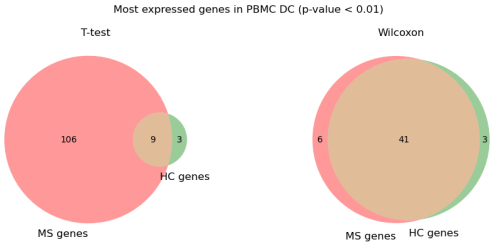
(d) Monocytes



(a) Plasma cells

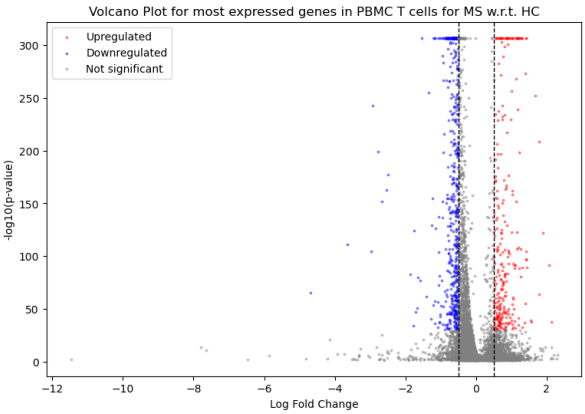


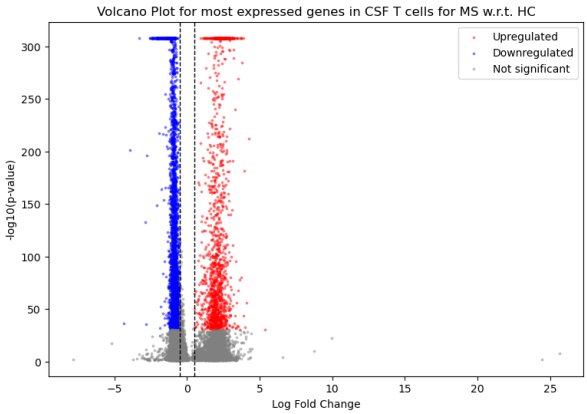
(b) Hematopoietic Stem Cell and Mltipotent Progenitor (HSC/MPP)



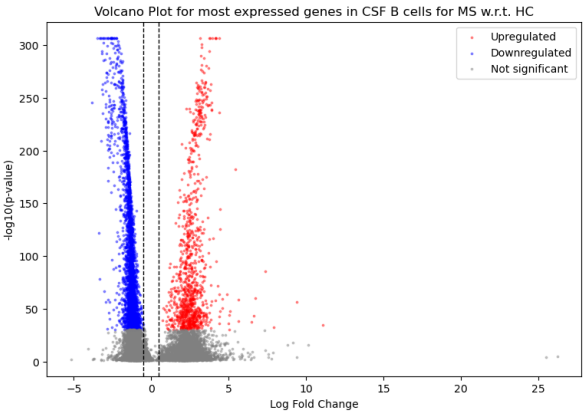
(c) Dendritic Cells (DCs)

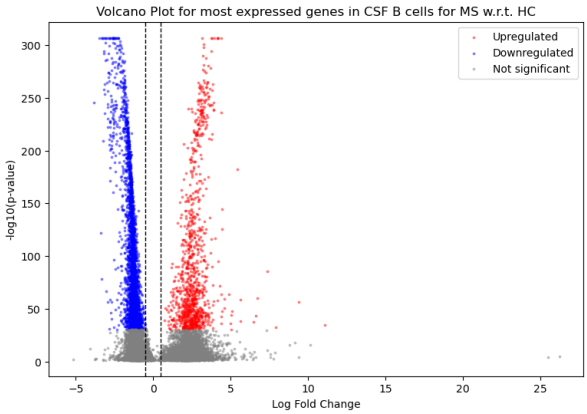
Figure 24: The Venn diagrams illustrate the most highly expressed genes identified for each cell type in both PBMC and CSF cells, with a p-value threshold of 0.01, comparing MS (Multiple Sclerosis) versus HC (Healthy Control). The diagrams compare the results of the t-test and Wilcoxon rank-sum test methods, highlighting the overlap and differences in gene expression identified by each method. This visualization aids in understanding the gene expression patterns specific to MS and HC across different cell types.



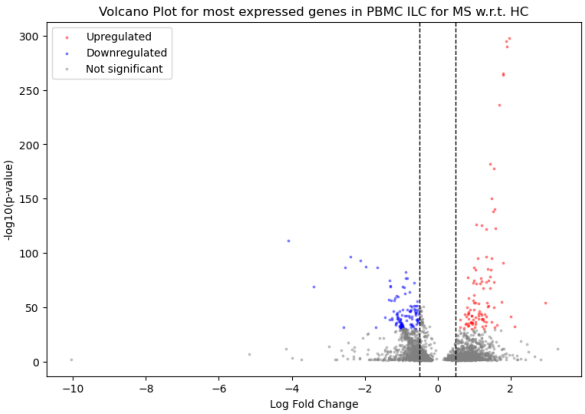


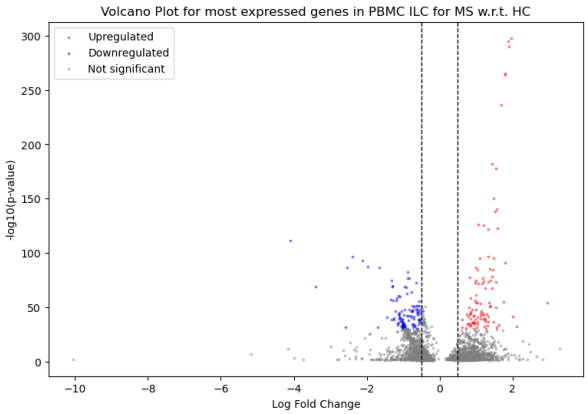
(a) T cells



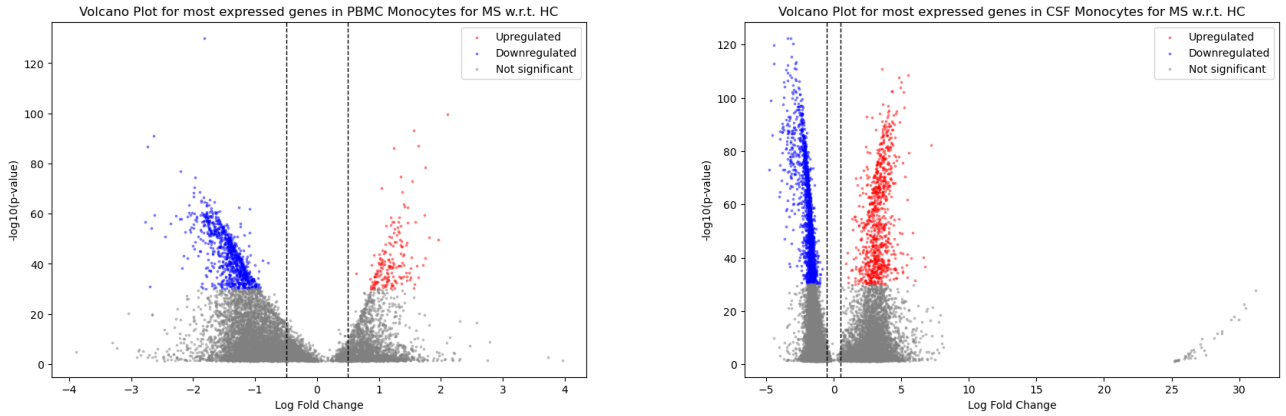


(b) B cells

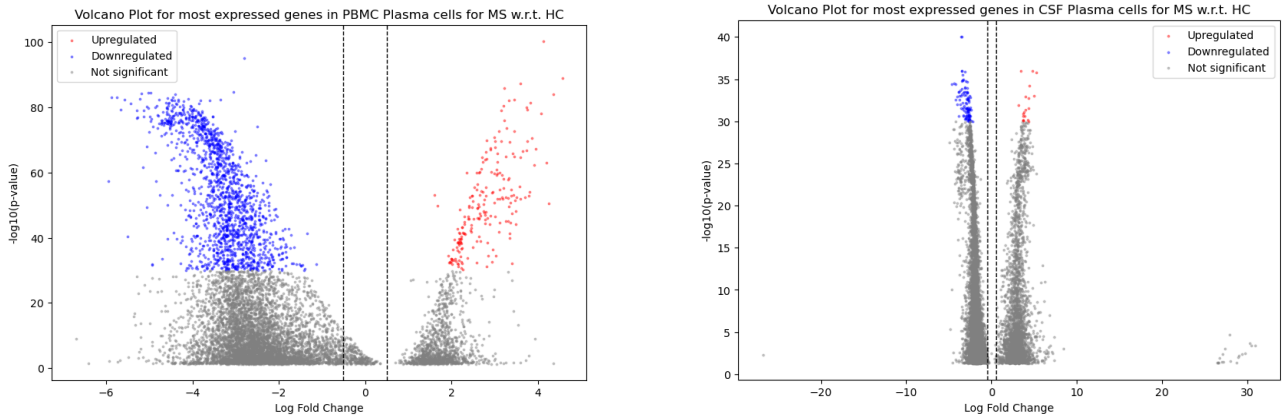




(c) Innate Lymphoid Cells (ILCs)



(a) Monocytes



(b) Plasma cells

Figure 26: Volcano plots visually illustrate genes that are upregulated (in red) and downregulated (in blue) in the cell populations from MS patients compared to those from HC samples. The criteria for determining significant genes include a threshold value for the adjusted p-value of 10^{-30} and a threshold for log fold change of 0.5. Genes shown in gray in the plot correspond to non-significant genes. The volcano plots are presented for each cell type in the following order, from top to bottom: T cells, B cells, ILC, Monocytes, and Plasma Cells. For each cell type, the volcano plot on the left represents PBMC cells, and the one on the right represents CSF cells.

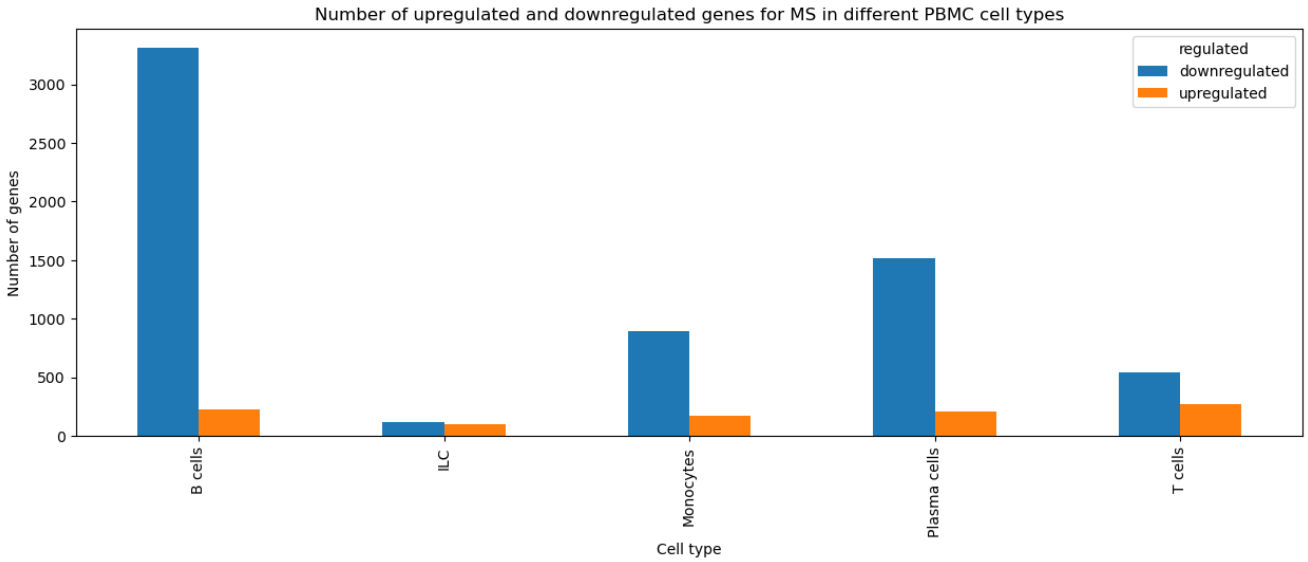


Figure 27: Barplot illustrating the number of upregulated and downregulated genes for each cell type in the PBMC (Peripheral Blood Mononuclear Cell) population. The data presented in the barplot is summarized in Table 5, indicating the count of genes for each cell type.

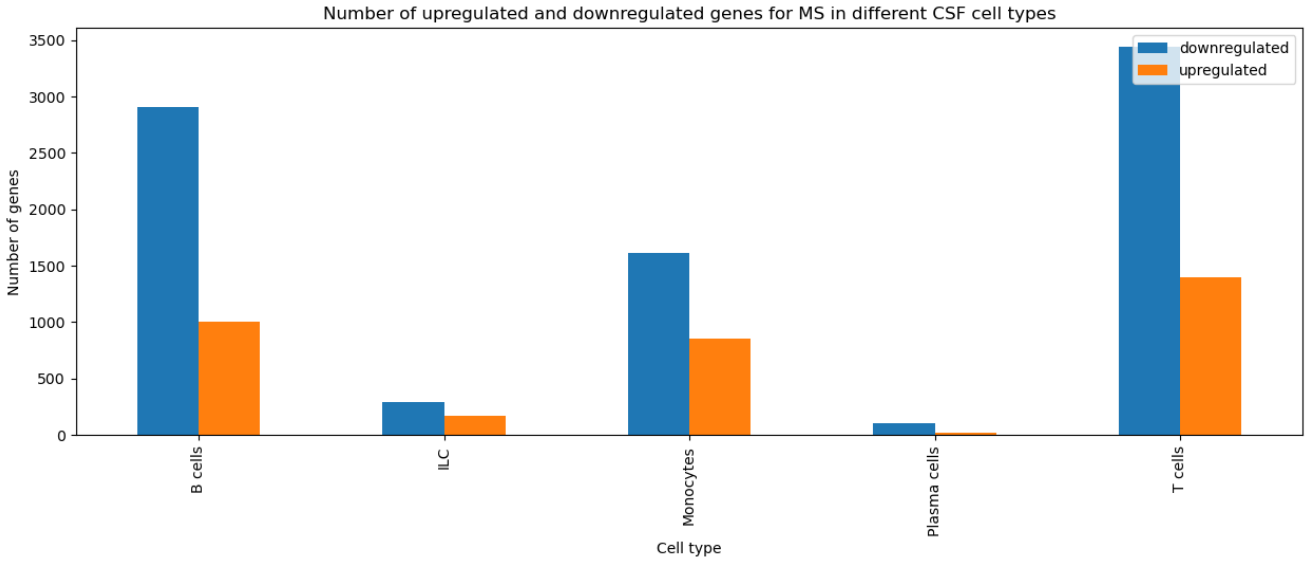


Figure 28: Barplot illustrating the number of upregulated and downregulated genes for each cell type in the CSF (Cerebrospinal Fluid) population. The data presented in the barplot is summarized in Table 5, indicating the count of genes for each cell type.

	List of Genes
Upregulated PBMC	-
Downregulated PBMC	RPL8, RPL38, CD48, HLA-DPB1, RPS13, RPL24, RPL12, FTL, UQCRB, FTH1, RPL27A, COX7C, RPS8, UBA52, NACA, SH3BGRL3, S100A6, RPL23, GAPDH, LGALS1, CCNI, RPS20, ATP5MC2, ATP5F1E, S100A4, BTF3, TMA7, RPL7, MYL12B, OAZ1, RPL31, HLA-DPA1, RPL29
Upregulated CSF	PCDHA2, CISD1, LINC00588, AJ003147.1, ETF1, AQP4-AS1, NOMO3, TEX52, AP001351.1, ESR2, AL359475.1, PCDHA4, LINC02417
Downregulated CSF	VAMP8, H3F3A, YBX1, MT-ND4, MT-ND5, HMGB1, SLC25A6, RHOA, TPM3, FTL, ITGB2, GABARAP, TLN1, SEPTIN7, PNRC1, RPL13A, CD99, TPI1, EMP3, ELOB, CDC42, RPS8, UQCRH, COX5B, SH3BGRL3, ARPC5, S100A6, MT-ND2, RPL23, GAPDH, COX6B1, S100A11, CYBA, ARPC1B, CFL1, TXNIP, COX8A, ARPC2, SERF2, IQGAP1, RPS10, RPS20, VIM, ATP5MC2, GNAI2, MTRNR2L12, SRP14, ATP5F1E, RPLP0, RPS17, S100A4, TMSB10, IFITM2, PTMA, GMFG, ITM2B, RPL7, OAZ1, CD63, ARPC3, RPL31, SET, HCST, RAC1

Table 11: Genes that are upregulated and downregulated in common across all cell types (T cells, B cells, Monocytes, Plasma cells, ILC) for PBMC and CSF. The tables demonstrate that 33 downregulated genes are common to all five cell types in PBMC. Furthermore, 13 upregulated genes and 64 downregulated genes are common to all five cell types in CSF, indicating potential universal mechanisms involved in MS pathology.

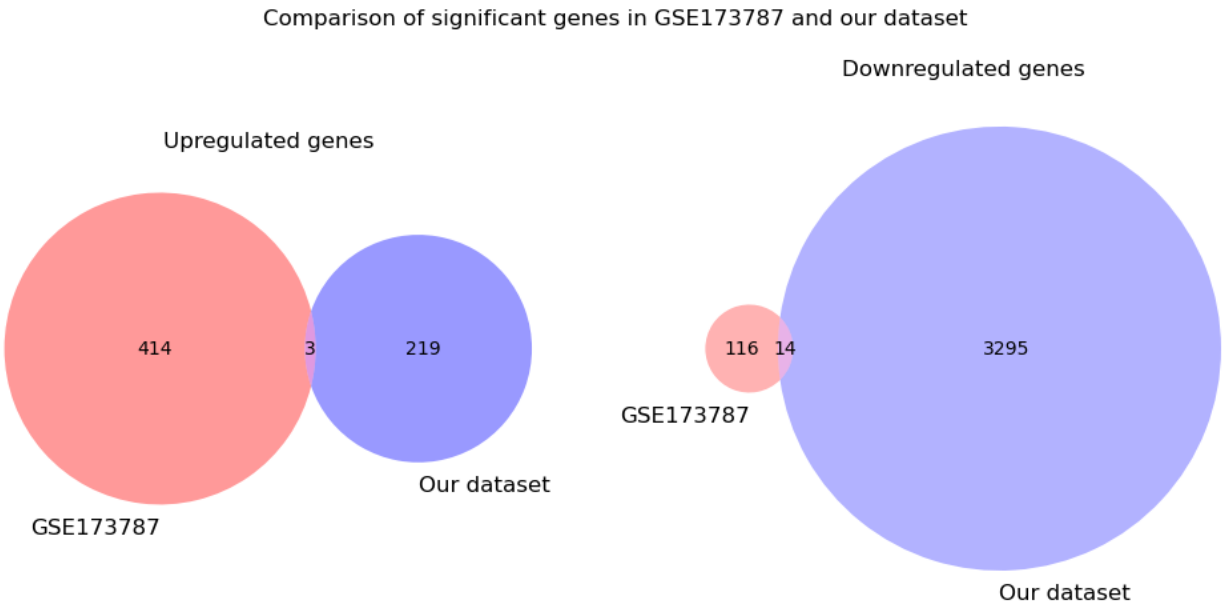


Figure 29: Venn diagram comparing the up and downregulated genes between the GSE173787 dataset and our transcriptomic dataset. Genes were determined using a threshold on the absolute value of the logFC value set to 0.5. In the GSE173787 dataset, 417 upregulated genes and 130 downregulated genes were identified for B cells. Conversely, our dataset revealed 222 upregulated genes and 3309 downregulated genes in the same cell type. Between the two datasets, only 3 upregulated genes and 14 downregulated genes were found to be shared. The shared upregulated genes are IGLC3, IGLC2, and IGKC, while the shared downregulated genes include CCDC28A, OGFRL1, BRI3, HSPB1, IFNGR1, MID1IP1, CSNK1E, EMC6, DUSP6, NDUFV2, DHRS3, CLEC2B, CMTM3, and BEX4.